APPLIED LOGISTIC REGRESSION ANALYSIS

SECOND EDITION

Scott Menard

Series: Quantitative Applications in the Social Sciences

106

a SAGE **UNIVERSITY** PAPER

Series/Number 07-106

APPLIED LOGISTIC REGRESSION ANALYSIS
Second Edition

SCOTT MENARD
Institute of Behavioral Science
University of Colorado

SAGE PUBLICATIONS
International Educational and Professional Publisher
Thousand Oaks London New Delhi

For information:

Sage Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

Sage Publications Ltd.
6 Bonhill Street
London EC2A 4PU
United Kingdom

Sage Publications India Pvt. Ltd.
M-32 Market
Greater Kailash I
New Delhi 110 048 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Menard, Scott W.
Applied logistic regression analysis / Scott Menard.–2nd ed.
p. cm. – (Quantitative applications in the social sciences
(Qass) ; v. 106)
Includes bibliographical references and index.
ISBN 0-7619-2208-3
1. Regression analysis. 2. Logistic distribution.
I. Title. II. Sage university papers series. Quantitative applications in the social sciences ; no. 106
QA278.2 .M46 2001
519.5′36–dc21 2001019806

This book is printed on acid-free paper.

05 06 07 10 9 8 7 6 5 4

Acquiring Editor:	C. Deborah Laughton
Editorial Assistant:	Veronica Novak
Production Editor:	Denise Santoyo
Typesetter:	Technical Typesetting Inc.

When citing a university paper, please use the proper form. Remember to cite the Sage University Paper series title and include paper number. One of the following formats can be adapted (depending on the style manual used):

(1) MENARD, SCOTT (2001) *Applied Logistic Regression Analysis.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-106. Thousand Oaks, CA: Sage.

OR

(2) Menard, Scott (2001). *Applied Logistic Regression Analysis.* (Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-106). Thousand Oaks, CA: Sage.

CONTENTS

SERIES EDITOR'S INTRODUCTION

The linear regression model provides a powerful device for organizing data analysis. Researchers focus on the explanation of a dependent variable, Y, as a function of multiple independent variables, from X_1 to X_k. Models are specified, variables are measured, and equations are estimated with ordinary least squares (OLS). All goes well if the classical linear regression assumptions are met. However, several assumptions are likely to be unmet if the dependent variable has only two or three response categories. In particular, with a dichotomous dependent variable, assumptions of homoskedasticity, linearity, and normality are violated, and OLS estimates are inefficient at best. The maximum likelihood estimation of a logistic regression overcomes this inefficiency, transforming $Y(1, 0)$ into a logit (log of the odds of falling into the "1" category).

Professor Menard fully explicates the estimation, interpretation, and diagnostics of such logistic regression models. The pedagogical portmanteau of his work is the parallel he continually draws to the linear regression model. The logistic counterparts to the OLS statistics—the R^2, the standard error of estimate, the t ratio, and the slope—are systematically presented. These parallels allow the analyst to step naturally from familiar terrain to new turf. Traditional regression diagnostics as well—the Studentized residual, leverage, dbeta—are included in an innovative logistic "protocol" for diagnostics. The last chapter dissects the problem of a polytomous dependent variable, with multiple ordered or unordered categories.

The discussion of the various computer packages is up-to-the minute and discriminating. For example, he notes that in SPSS 10 the NOMREG routine is good for nominal dependent variables, whereas for ordered dependent variables, SAS LOGISTIC is preferred. Attention to current computer software is part of the changes made from the first edition of this monograph. Other changes include a comprehensive evaluation of the many different goodness-of-fit measures. Dr. Menard makes a convincing case for the use of R_L^2, at least if the goal is direct comparison to the OLS R^2. He also adds new material on grouped data, predictive efficiency, and risk ratios.

The Qualitative Applications in the Social Sciences series has published many papers on classical linear regression (see Lewis-Beck, *Applied Regression*, Vol. 22; Achen, *Interpreting and Using Regression*, Vol. 29; Berry and Feldman, *Multiple Regression in Practice*, Vol. 50; Schroeder, Sjoquist, and Stephan, *Understanding Regression Analysis*, Vol. 57; Fox, *Regression Diagnostics*, Vol. 79; Berry, *Understanding Regression Assumptions*, Vol. 92; and Hardy, *Regression With Dummy Variables*, Vol. 93). The voluminous output is justified by the dominance of the linear regression paradigm. In observational research work, this paradigm increasingly is bumping up against the reality of dependent variables that are less than continuous, less than interval. Hence the heightened attention to the logistic regression alternative. The series first published DeMaris, *Logit Modeling*, Vol. 86, followed by Menard, *Applied Logistic Regression Analysis, 1st ed.*, Vol. 106, and Pampel, *Logistic Regression: A Primer*, Vol. 132. The last mentioned monograph provides a basic introduction to the technique. The monograph at hand goes beyond that introduction, attending to the most contemporary of complex issues and mechanics. For the social scientist who wishes to be au courant regarding this rapidly evolving topic, the Menard second edition is a must.

—*Michael S. Lewis-Beck*
Series Editor

AUTHOR'S INTRODUCTION TO THE SECOND EDITION

The last sentence before the Notes section in the previous edition of this monograph was, "One can hope that many of the 'kludges' for making logistic regression analysis work with existing software ...will become obsolete as the software available for logistic regression analysis is expanded and improved." This second edition was written because that hope has been at least partially fulfilled. There have been some relatively minor changes in SAS, including the addition of the Cox–Snell and Nagelkerke pseudo-R^2 measures, plus an optional component, SAS Display Manager, a windowing shell. There are also two new instructional manuals devoted to logistic regression analysis (Allison, 1999; SAS, 1995) that have been published since the first edition of this monograph. SPSS PC+, a command-driven package, which was illustrated in the first edition, has been supplanted by SPSS 10, which is highly integrated with the Windows 95/98 environment. Also, there are two new SPSS routines, NOMREG (nominal regression) for polytomous nominal logistic regression and PLUM (polytomous logit universal model) for ordinal logistic regression and related models (Norusis, 1999; SPSS, 1999a). Changes from the first edition include:

- More detailed consideration of grouped as opposed to casewise data throughout the monograph
- An updated discussion of the properties and appropriate use of goodness-of-fit measures, R^2 analogues, and indices of predictive efficiency (Chapter 2);
- Discussion of the misuse of odds ratios to represent risk ratios (Chapter 3)
- Discussion of overdispersion and underdispersion for grouped data (Chapter 4)
- Updated coverage of unordered and ordered polytomous logistic regression models; some material that is no longer necessary for working around the limitations of earlier versions of the software has been dropped (Chapter 5).

The focus in this second edition, as in the first, is on logistic regression models for individual level data, but aggregate or grouped data, with multiple cases for each possible combination of values of the predictors, are considered in more detail. As in the first edition, examples using SAS and SPSS software are provided. Finally, observant readers informed me about places in the previous edition where there were errors or where the clarity of presentation could be improved. For their comments, questions, and constructive criticisms, I thank the anonymous reviewers of the first edition, Alfred De Maris, who reviewed the present edition, and also Dennis Fisher, Tom Knapp, Michael Lewis-Beck, Fred Pampel, Hidetoshi Saito, Dan Waschbusch, Susan White, and, especially, David Nichols of SPSS for his detailed comments. I would also like to absolve all of them of blame for any errors, new or old, in the present edition.

APPLIED LOGISTIC REGRESSION ANALYSIS
Second Edition

SCOTT MENARD
Institute of Behavioral Science
University of Colorado

1. LINEAR REGRESSION AND THE LOGISTIC REGRESSION MODEL

In linear regression analysis, it is possible to test whether two variables are linearly related and to calculate the strength of the linear relationship if the relationship between the variables can be described by an equation of the form $Y = \alpha + \beta X$, where Y is the variable being predicted (the dependent, criterion, outcome, or endogenous variable), X is a variable whose values are being used to predict Y (the independent, exogenous, or predictor variable),[1] and α and β are population parameters to be estimated. The parameter α, called the *intercept*, represents the value of Y when $X = 0$. The parameter β represents the change in Y associated with a one-unit increase in X or the *slope* of the line that provides the best linear estimate of Y from X. In *multiple regression*, there are several predictor variables. If k denotes the number of independent variables, the equation becomes $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$ and $\beta_1, \beta_2, \ldots, \beta_k$ are called *partial* slope coefficients, reflecting the fact that any one of the k predictor variables $X_1, X_2, \ldots, X_k$ provides only a partial explanation or prediction for the value of Y. The equation is sometimes written in a form that explicitly recognizes that prediction of Y from X may be imprecise, $Y = \alpha + \beta X + \epsilon$, or for several predictors, $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$, where ϵ is the error term, a random variable that represents the error in predicting Y from X. For an individual case j, $Y_j = \alpha_j + \beta X_j + \epsilon_j$ or $Y_j = \alpha_j + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj} + \epsilon_j$, and the subscript j indicates that the equation predicts values for specific cases, indexed by j ($j = 1$

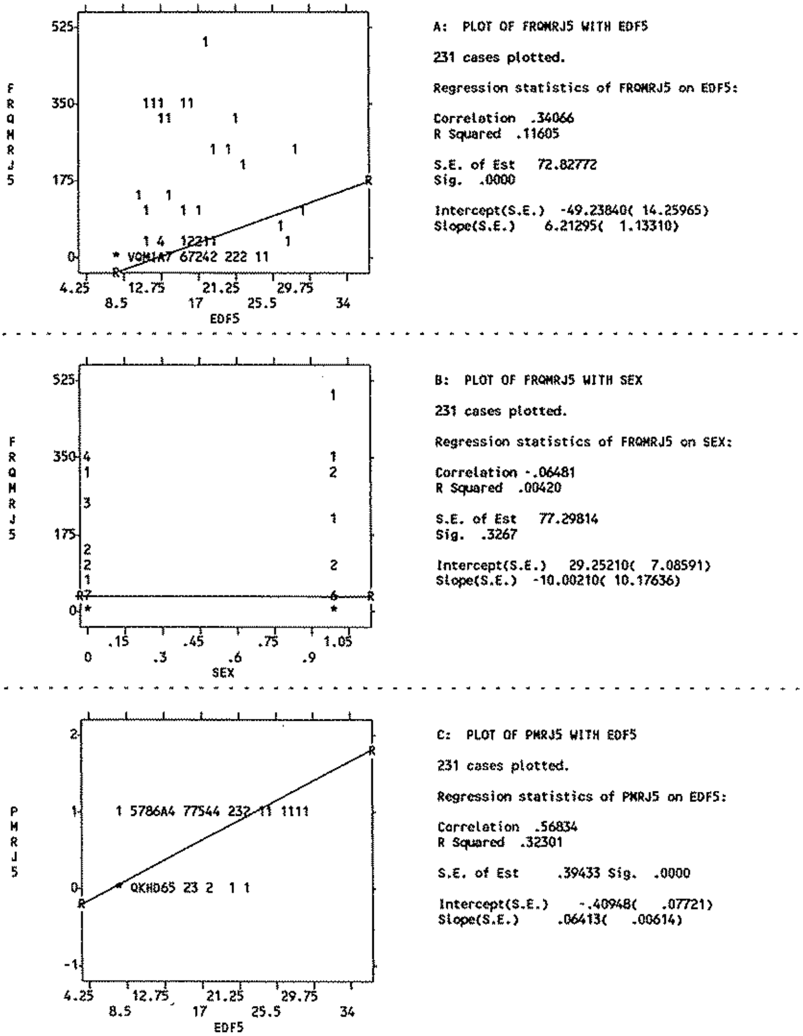
for the first case, $j = 2$ for the second case, etc.). Y_j, X_{1j}, X_{kj}, etc. refer to specific values of the dependent and independent variables. This last equation is used to calculate the value of Y for a particular case j, rather than describing the relationship among the variables for all of the cases in the sample or the population.

Estimates of the intercept α and the regression coefficients β (or β_1, β_2, ..., β_k) are obtained mathematically using the method of ordinary least squares (OLS) estimation, which is discussed in many introductory statistics texts (for example, Agresti and Finlay, 1997; Bohrnstedt and Knoke, 1994). These estimates produce the equation $\widehat{Y} = a + bX$ or, in the case of several predictors, $\widehat{Y} = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k$, where $\widehat{Y}$ is the value of Y predicted by the linear regression equation, a is the OLS estimate of the intercept α, and b (or b_1, b_2, ..., b_k) is the OLS estimate for the slope β (or the partial slopes β_1, β_2, ..., β_k). Residuals for each case e_j are equal to $(Y_j - \widehat{Y}_j)$, where $\widehat{Y}_j$ is the estimated value of Y_j for case j. For bivariate regression, the residuals can be visually or geometrically represented by the vertical distance between each point in a bivariate scatterplot and the regression line. For multiple regression, visual representation is much more difficult because it requires several dimensions.

An example of a bivariate regression model is given in Figure 1.1. In part A of Figure 1.1, the dependent variable is FRQMRJ5, the annual frequency of self-reported marijuana use ("How many times in the last year have you smoked marijuana?"), and the independent variable is EDF5, an index of exposure to delinquent friends, for 16-year-old respondents interviewed in 1980 in the fifth wave of a national household survey.[2] The exposure to delinquent friends scale is the sum of the answers to eight questions about how many of the respondent's friends are involved in different types of delinquent behavior (theft, assault, drug use). The responses to individual items range from 1 (none of my friends) to 5 (all of my friends), resulting in a possible range from 8 to 40 for EDF5. From part A of Figure 1.1, there appears to be a positive relationship between exposure to delinquent friends and marijuana use, described by the equation

$$(\text{FRQMRJ5}) = -49.2 + 6.2(\text{EDF5}).$$

In other words, for every one-unit increase in the index of exposure to delinquent friends, frequency of marijuana use increases by about six times per year, or about once every two months. The coefficient



```
A:  PLOT OF FRQMRJ5 WITH EDF5

231 cases plotted.

Regression statistics of FRQMRJ5 on EDF5:

Correlation  .34066
R Squared  .11605

S.E. of Est   72.82772
Sig.  .0000

Intercept(S.E.)  -49.23840( 14.25965)
Slope(S.E.)    6.21295(  1.13310)
```

```
B:  PLOT OF FRQMRJ5 WITH SEX

231 cases plotted.

Regression statistics of FRQMRJ5 on SEX:

Correlation -.06481
R Squared  .00420

S.E. of Est   77.29814
Sig.  .3267

Intercept(S.E.)   29.25210(  7.08591)
Slope(S.E.)  -10.00210( 10.17636)
```

```
C:  PLOT OF PMRJ5 WITH EDF5

231 cases plotted.

Regression statistics of PMRJ5 on EDF5:

Correlation  .56834
R Squared  .32301

S.E. of Est     .39433 Sig.  .0000

Intercept(S.E.)    -.40948(   .07721)
Slope(S.E.)      .06413(   .00614)
```

Figure 1.1. Bivariate Regression Plots

of determination (R^2) indicates how much better we can predict the dependent variable from the independent variable than we could predict the dependent variable without information about the independent variable. Without information about the independent variable, we would use the mean frequency of marijuana use as our prediction for all respondents. By knowing the value of exposure to delinquent friends, however, we can base our prediction on the value of EDF5 and the relationship, represented by the regression equation, between FRQMRJ5 and EDF5. Using the regression equation reduces the sum of the squared errors of prediction, $\sum e_j^2 = \sum(\widehat{Y}_j - Y_j)^2$, by $R^2 = .116$ or about 12%.

It is necessary when interpreting the results to consider the actual values of the dependent and independent variables. The intercept indicates that for an individual with 0 as the value of exposure to delinquent friends, the frequency of marijuana use would be negative. This seemingly impossible result occurs because exposure, as already noted, is measured on a scale that ranges from a minimum of 8 (no exposure at all; not one friend is involved in any of 8 delinquent activities) to 40 (extensive exposure; all friends are involved in all 8 delinquent activities). Thus, for individuals with the minimum possible exposure to delinquent friends (a value of 8, representing no exposure), the expected frequency of marijuana use is $-49.2+6.2(8) = 0.4$, which is close to 0, but indicates that even some individuals with no exposure to delinquency may use marijuana at least occasionally. The maximum value of EDF5 in this sample is 29, which corresponds to an expected frequency of marijuana use equal to $-49.2+6.2(29) = 130.6$ or use approximately every 3 days. This result makes sense substantively, in terms of real-world behavior, as well as statistically, in terms of the regression equation.

1.1. Regression Assumptions

To use the OLS method to estimate and make inferences about the coefficients in linear regression analysis, a number of assumptions must be satisfied (Lewis-Beck, 1980, pp. 26–47; Berry & Feldman, 1985; Berry, 1993). Specific assumptions include the following:

1. *Measurement:* All independent variables are interval, ratio, or dichotomous, and the dependent variable is continuous, unbounded, and measured on an interval or ratio scale. All variables are measured without error.[3]

2. *Specification:* (a) All relevant predictors of the dependent variable are included in the analysis, (b) no irrelevant predictors of the dependent variable are included in the analysis, and (c) the *form* of the relationship (allowing for transformations of dependent or independent variables) is linear.
3. *Expected value of error:* The expected value of the error, ϵ, is 0.
4. *Homoscedasticity:* The variance of the error term, ϵ, is the same or constant for all values of the independent variables.
5. *Normality of errors:* The errors are normally distributed for each set of values of the independent variables.
6. *No autocorrelation:* There is no correlation among the error terms produced by different values of the independent variables. Mathematically, $E(\epsilon_i, \epsilon_j) = 0$.
7. *No correlation between the error terms and the independent variables:* The error terms are uncorrelated with the independent variables. Mathematically, $E(\epsilon_j, X_j) = 0$.
8. *Absence of perfect multicollinearity:* For multiple regression, none of the independent variables is a perfect linear combination of the other independent variables. Mathematically, for any i, $R_i^2 < 1$, where R_i^2 is the variance in the *independent* variable X_i that is explained by all of the other independent variables $X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k$. If there is only one predictor, multicollinearity is not an issue.

1.1.1. Violations of the Measurement Assumption: Dichotomous Variables in Linear Regression

The linear regression model can be extended easily to accommodate dichotomous predictors, including sets of dummy variables (Lewis-Beck, 1980, pp. 66–71; Berry & Feldman, 1985, pp. 64–75; Hardy, 1993). An example is presented in part B of Figure 1.1. Here, the dependent variable is again self-reported annual frequency of marijuana use, but the independent variable this time is sex or gender (coded 0=female, 1=male). The regression equation is

$$\text{FRQMRJ5} = 29.3 - 10.0(\text{SEX}).$$

The resulting diagram consists of two columns of values for frequency of marijuana use: one represents females and one represents males. With a dichotomous predictor, coded 0–1, the intercept and the slope have a special interpretation. It is still true that the intercept is the predicted value of the dependent variable when the independent variable is 0 (substantively, when the respondent is female), but with only two groups, the intercept now is the *mean* frequency of marijuana use for the group coded as 0 (females). The slope is still the change in the dependent variable associated with a one-unit change in the independent variable, but with only two categories, that value becomes the *difference in the means* between the first (female) and second (male) groups. The sum of the slope and the intercept, 29.3 − 10.0 = 19.3, is therefore the *mean* frequency of marijuana use for the second group (males). As indicated in part B of Figure 1.1, females report a higher (yes, higher) frequency of marijuana use than males, but the difference is not statistically significant (as indicated by Sig. = .3267). In part B of Figure 1.1, the regression line is simply the line that connects the mean frequency of marijuana use for females and the mean frequency of marijuana use for males, that is, the *conditional means*[4] of marijuana use for females and males, respectively. The predicted values of Y over the observed range of X lie well within the observed (and possible) values of Y. Again, the results make substantive as well as statistical sense.

When the *dependent* variable is dichotomous, the interpretation of the regression equation is not as straightforward. In part C of Figure 1.1, the independent variable is again exposure to delinquent friends, but now the dependent variable is the *prevalence* of marijuana use: whether (yes = 1 or no = 0) the individual used marijuana at all during the past year. In part C of Figure 1.1, with a dichotomous dependent variable, there are two rows (rather than columns, as in part B). The linear regression model with a dichotomous dependent variable, coded 0–1, is called a *linear probability model* (Agresti, 1990, p. 84; Aldrich & Nelson, 1984). The equation from part C of Figure 1.1 is

$$\text{PMRJ5} = -.41 + .064(\text{EDF5}).$$

When there is a dichotomous dependent variable, the mean of the variable is a function of the probability[5] that a case will fall into the higher of the two categories for the variable. Coding the values of the variable as 0 and 1 produces the result that the mean of the variable is the proportion of cases in the higher of the two categories of the variable, and the predicted value of the dependent variable (the conditional mean, given the value of X and the assumption that

X and Y are linearly related) can be interpreted as the *predicted probability* that a case falls into the higher of the two categories on the dependent variable, given its value on the independent variable. Ideally, we would like the predicted probability to lie between 0 and 1, because a probability cannot be less than 0 or more than 1.

As is evident from part C of Figure 1.1, the predicted values for the dependent variable may be higher or lower than the possible values of the dependent variable. For the minimum value of EDF5 (EDF5 = 8), the predicted prevalence of marijuana use (i.e., the predicted probability of marijuana use) is $-.41 + .06(8) = .10$, a reasonable result, but for the maximum value of EDF5 (EDF5 = 29), the predicted probability of marijuana use becomes $-.41 + .064(29) = 1.45$, or an impossibly high probability of about $1\frac{1}{2}$. In addition, the variability of the residuals will depend on the size of the independent variable (Schroeder, Sjoquist, & Stephan, 1986, pp. 79–80; Aldrich & Nelson, 1984, p. 13). This condition, called *heteroscedasticity*, implies that the estimates for the regression coefficients, although they are unbiased (not systematically too high or too low), will not be the best estimates in the sense of having a small standard error. There is also a systematic pattern to the values of the residuals that depends on the value of X. For values of X greater than 23.5 in part C of Figure 1.1, all of the residuals will be negative because $\widehat{Y}_j$ will be greater than Y_j (because for X greater than 23.5, $\widehat{Y}_j$ is greater than 1, but Y_j is less than or equal to 1). Also, residuals will not be normally distributed (Schroeder et al., 1986, p. 80) and sampling variances will not be correctly estimated (Aldrich & Nelson, 1984, pp. 13–14); therefore, the results of hypothesis testing or construction of confidence intervals for the regression coefficients will not be valid.

1.1.2. Nonlinearity, Conditional Means, and Conditional Probabilities

For continuous dependent variables, $\widehat{Y}$, the regression estimate of Y, may be thought of as an *estimate* of the conditional mean of Y for a particular value of X, *given that the relationship between X and Y is linear*. In bivariate regression, for continuous independent variables, the estimated value of Y may not be exactly equal to the mean value of Y for those cases, because the conditional means of Y for different values of X may not lie exactly on a straight line. For a dichotomous predictor variable, the regression line will pass exactly through the conditional means of Y for each of the two categories of X. If the conditional means of FRQMRJ5 are plotted against the dichotomous predictor SEX, the plot consists of two points (remember, the cases are aggregated by the value of the independent variable): the conditional means of Y for males and females. The simplest, most parsimonious description of this plot is a straight line between the two conditional means, and the linear regression model appears to work well.

The inherent nonlinearity of relationships that involve dichotomous dependent variables is illustrated in Figure 1.2. In Figure 1.2, the observed conditional mean of PMRJ5, the *prevalence* of marijuana use, is plotted for each value of the independent variable EDF5. The observed conditional mean is symbolized by the letter "C." Since PMRJ5 is coded as either 0 or 1, the conditional means represent averages of 0s and 1s, and are interpretable as conditional probabilities. Figure 1.2 is therefore a plot of *probabilities* that PMRJ5 = 1 for different values of EDF5. All of the *observed* values of Y lie between the two vertical lines at 0 and 1, respectively, in Figure 1.2. *Predicted* probabilities, however, can, in principle, be infinitely large or small if we use the linear probability model.

The plot of *observed* conditional probabilities (C) in Figure 1.2 is overlaid with the plot of *predicted* conditional probabilities based on the regression equation (R) in part C of Figure 1.1. For values of EDF5 greater than 23.5, the observed value of the conditional mean prevalence of marijuana use stops increasing and levels off at PMRJ5 = 1. The predicted values from the regression equation, however, continue to increase past the value of 1 for PMRJ5, to a maximum of 1.45, and the error of prediction increases as EDF5 increases from 23.5 to its maximum of 29.

Two points need to be made about Figure 1.2. First, although a linear model appears to be potentially appropriate for a continuous dependent variable, regardless of whether the independent variables are continuous or dichotomous, it is evident that a nonlinear model is better suited to the analysis of the dichotomous variable PMRJ5. In general, for very high values of X (or very low values, if the relationship is negative), the conditional probability that $Y = 1$ will be so close to 1 that it should change little with further increases in X. This is the situation illustrated in Figure 1.2. It is also the case that for very low values of X (or very high values if the relationship is negative), the conditional probability that $Y = 1$ will be so close to 0 that it should change little with further decreases in X. The curve that represents the relationship between X and Y should, therefore,

```
         ++----+----+----+----+----+----+----+----+----+----+----+----+----++
      1.6+                                                                  +
         |                                                                  |
         |                                                                  |
         |                                                                  |
         |                                                                R |
      1.4+                                                              R   +
         |                                                                  |
         |                                                            R     |
         |                                                                  |
         |                                                          R       |
      1.2+                                                        R         +
M        |                                                                  |
P        |                                                     R            |
M        |                                                                  |
R        |                                                                  |
J       1+                                      C   C    C R   C  C C C C C +
5        |                                                                  |
         |                                               R                  |
M        |                                            R                     |
R        |                                                                  |
P      .8+                                          R                       +
E        |                                 C      R                         |
P        |                                    C                             |
M        |                                      R C   C    C                |
J        |                             C                                    |
5      .6+                                    R                             +
         |                                 R                                |
         |                                                                  |
         |                               R                                  |
         |                             R C                                  |
       .4+                                                                  +
         |                           R                                      |
         |                        C  C                                      |
         |                        R                                         |
         |                      $                                           |
       .2+                                                                  +
         |                    $                                             |
         |                  R                                               |
         |                                                                  |
         |                                                                  |
        0+                  C                                               +
         ++----+----+----+----+----+----+----+----+----+----+----+----+----++
              2.25      6.75     11.25     15.75     20.25     24.75     29.25
           0       4.5         9      13.5        18      22.5        27

                                     EDF5
```

Figure 1.2. Conditional Probabilities Observed (C) and Predicted by Linear Regression (R)
C: Observed Mean Prevalence of Marijuana Use (MPMRJ5) WITH EDF5 (Exposure to Delinquent Friends). R: Linear Regression Prediction of Prevalence of Marijuana Use (MRPEPMJ5) WITH EDF5. $: Multiple occurrence (Linear Regression prediction and observed value coincide). 21 cases.

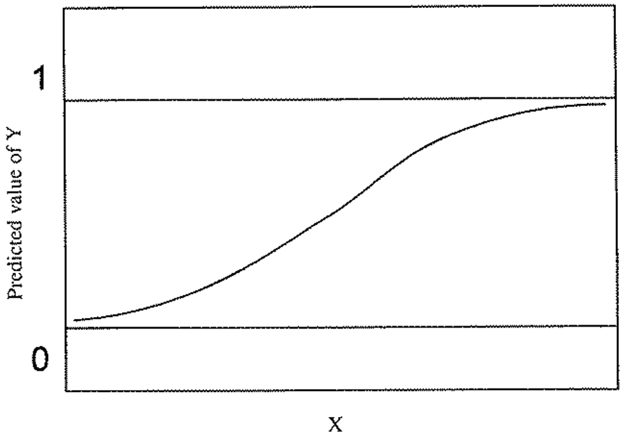


Figure 1.3. Logistic Curve Model for a Dichotomous Dependent Variable

be very shallow, with a slope close to 0, for very high and very low values of X if X can, in principle, become indefinitely large or indefinitely small. If X and Y are related, then between the very high and very low values of X, the slope of the curve will be steeper, that is, significantly different from 0. The general pattern is that of an "S curve" as depicted in Figure 1.3.

Second, for prevalence data, the observed conditional mean of Y is equal to the observed conditional probability that $Y = 1$, and the predicted value of Y is equal to the predicted conditional probability that $Y = 1$. The actual values used to identify the two categories of Y are arbitrary, a matter of convenience. They may be 0 and 1, for example, or 2 and 3 (in which case the predicted values of Y are equal to 2 plus the conditional probability that $Y = 3$, still a function of the conditional probability that Y has the higher of its two values for a given value of X). What is substantively important is not the numerical value of Y, but the probability that Y has one or the other of its two possible values, and the extent to which that probability depends on one or more independent variables.

The distinction between the arbitrary numerical value of Y, upon which OLS parameter estimates are based, and the probability that Y has one or the other of its two possible values is problematic for OLS linear regression and leads us to consider alternative methods for estimating parameters to describe the relationship between X and Y.

First, however, we address the issue of nonlinearity. For continuous independent and dependent variables, the presence of nonlinearity in the relationship between X and Y may sometimes be addressed by the use of nonlinear transformations of dependent or independent variables (Berry & Feldman, 1985). Similar techniques play a part in estimating relationships that involve dichotomous dependent variables.

1.2. Nonlinear Relationships and Variable Transformations

When a relationship appears to be nonlinear, it is possible to transform either the dependent variable or one or more of the independent variables so that the *substantive* relationship remains nonlinear, but the *form* of the relationship is linear and can, therefore, be analyzed using OLS estimation. Another way to say that a relationship is substantively nonlinear but formally linear is to say that the relationship is *nonlinear in terms of its variables, but linear in terms of its parameters* (Berry & Feldman, 1985, p. 53). Examples of using variable transformations to achieve a linear form for the relationship are given in Berry and Feldman (1985, pp. 55–72) and Lewis-Beck (1980, pp. 43–47).

In Figure 1.2, there was some evidence of nonlinearity in the relationship between frequency of marijuana use and exposure to delinquent friends. One possible transformation that could be used to model this nonlinearity is a logarithmic transformation[6] of the dependent variable FRQMRJ5. This is done by adding 1 to FRQMRJ5 and then taking the natural logarithm. (Adding 1 is necessary to avoid taking the natural logarithm of 0, which is undefined.) The regression equation then has the form $\ln(Y+1) = \alpha + \beta X$ or, equivalently, $(Y+1) = e^{\alpha+\beta X}$ or $Y = e^{\alpha+\beta X} - 1$, where $e = 2.72$ is the base of the natural logarithm. Specifically, for prevalence of marijuana use and exposure to delinquent friends,

$$\ln(\text{FRQMRJ5} + 1) = -1.7 + .23(\text{EDF5}), \qquad R^2 = .32.$$

Comparing the results of the model using the logarithmic transformation with the untransformed model in part A of Figure 1.1, it is evident that the slope is still positive, but the numerical value of the slope has changed (because the units in which the dependent variable is measured have changed from frequency to log frequency).

The coefficient of determination for the transformed equation is also larger (.32 instead of .12), reflecting a better fit of the linear regression model when the dependent variable is transformed. This is evidence (not conclusive proof, just evidence) that the relationship between frequency of marijuana use and exposure to delinquent friends is substantively nonlinear. A similar result occurs for the relationship between the dichotomous predictor SEX and frequency of marijuana use. With the logarithmic transformation of the dependent variable, the explained variance increases (from a puny .004 to an unimpressive .028), and the relationship between gender and frequency of marijuana use is statistically significant ($p = .011$) in the transformed equation. It appears that the relationship between frequency of marijuana use and both of the predictors considered so far is substantively nonlinear, but we are still able to use a formal linear model to describe those relationships and we can still use OLS to estimate the parameters of the model.

1.3. Probabilities, Odds, Odds Ratios, and the Logit Transformation for Dichotomous Dependent Variables

As noted earlier, for a dichotomous dependent variable, the numerical value of the variable is arbitrary, a matter of convenience, and is not intrinsically interesting. What is intrinsically interesting is whether the classification of cases into one or the other of the categories of the dependent variable can be predicted by the independent variable. Instead of trying to predict the arbitrary value associated with a category, it may be useful to reconceptualize the problem as an attempt to predict the probability that a case will be classified into one as opposed to the other of the two categories of the dependent variable. Because the probability of being classified into the first or lower-valued category, $\mathrm{P}(Y = 0)$, is equal to 1 minus the probability of being classified into the second or higher-valued category, $\mathrm{P}(Y = 1)$, if we know one probability, we know the other: $\mathrm{P}(Y = 0) = 1 - \mathrm{P}(Y = 1)$.

We could try to model the probability that $Y = 1$ as $\mathrm{P}(Y = 1) = \alpha + \beta X$, but we would again run into the problem that although observed values of $\mathrm{P}(Y = 1)$ must lie between 0 and 1, predicted values may be less than 0 or greater than 1. A step toward solving this problem would be to replace the *probability* that $Y = 1$ with the *odds* that $Y = 1$. The *odds* that $Y = 1$, written $\mathrm{odds}(Y = 1)$, is the ratio of the probability that $Y = 1$ to the probability that $Y \neq 1$. The odds that

$Y = 1$ is equal to $\mathrm{P}(Y = 1)/[1 - \mathrm{P}(Y = 1)]$. Unlike $\mathrm{P}(Y = 1)$, the odds has no fixed maximum value, but like the probability, it has a minimum value of 0.

One further transformation of the odds produces a variable that varies, in principle, from negative infinity to positive infinity. The *natural logarithm of the odds*, $\ln\{\mathrm{P}(Y = 1)/[1 - \mathrm{P}(Y = 1)]\}$, is called the *logit* of Y. The logit of Y, written logit(Y), becomes negative and increasingly large in absolute value as the odds decrease from 1 toward 0, and becomes increasingly large in the positive direction as the odds increase from 1 to infinity. If we use the natural logarithm of the odds that $Y = 1$ as our dependent variable, we no longer face the problem that the estimated probability may exceed the maximum or minimum possible values for the probability. The equation for the relationship between the dependent variable and the independent variables then becomes

$$\mathrm{logit}(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k. \qquad [1.1]$$

We can convert logit(Y) back to the odds by *exponentiation*, calculating [odds that $Y = 1$] $= e^{\mathrm{logit}(Y)}$. This results in the equation

$$\begin{aligned}\mathrm{odds}(Y = 1) &= e^{\ln[\mathrm{odds}(Y=1)]} \\ &= e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}\end{aligned} \qquad [1.2]$$

and a change of one unit in X multiplies the odds by e^{β}. We can then convert the odds back to the probability that $(Y = 1)$ by the formula $\mathrm{P}(Y = 1) =$ [odds that $Y = 1$]/[1 + odds that $Y = 1$], that is, the probability that $Y = 1$ is equal to the odds that $Y = 1$ divided by 1 plus the odds that $Y = 1$. This produces the equation

$$\mathrm{P}(Y = 1) = \frac{e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}{1 + e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}. \qquad [1.3]$$

It is important to understand that the probability, the odds, and the logit are three different ways to express exactly the same thing. Of the three measures, the probability or the odds is probably the most easily understood. Mathematically, however, the logit form of the probability best helps us to analyze dichotomous dependent variables. Just as we took the natural logarithm of the continuous dependent variable (frequency of marijuana use) to correct for the nonlinearity in the relationship between frequency of marijuana use and exposure to delinquent friends, we can also take the logit of the dichotomous dependent variable (prevalence of marijuana use) to correct for the nonlinearity in the relationship between prevalence of marijuana use and exposure to delinquent friends.

For any given case, logit(Y) $= \pm\infty$. This ensures that the probabilities estimated for the probability form of the model (Equation 1.3) will not be less than 0 or greater than 1, but it also means that because the linear form of the model (Equation 1.1) has infinitely large or small values of the dependent variable, OLS cannot be used to estimate the parameters. Instead, *maximum likelihood* techniques are used to maximize the value of a function, the *log-likelihood* function, which indicates how likely it is to obtain the observed values of Y, given the values of the independent variables and parameters $\alpha, \beta_1, \ldots, \beta_k$. Unlike OLS, which is able to solve directly for the parameters, the solution for the logistic regression model is found by beginning with a tentative solution, revising it slightly to see if it can be improved, and repeating the process until the change in the likelihood function from one step of the process to another is negligible. This process of repeated estimation, testing, and reestimation is called *iteration*, and the process of obtaining a solution from repeated estimation is called an *iterative* process. When the change in the likelihood function from one step to another becomes negligible, the solution is said to *converge*. All of this is done by means of computer-implemented numerical algorithms designed to search for and identify the best set of parameters to maximize the log-likelihood function. When the assumptions of OLS regression are met, however, *the OLS estimates for the linear regression coefficients are identical to the estimates that would be obtained using maximum likelihood estimation* (Eliason, 1993, pp. 13–18). OLS estimation is in this sense a special case of maximum likelihood estimation, one in which it is possible to calculate a solution directly without iteration.

1.4. Logistic Regression: A First Look

Part C of Figure 1.1 showed the results of an OLS linear regression analysis of the relationship between the prevalence of marijuana use (PMRJ5) and exposure to delinquent friends (EDF5). Figure 1.4 presents the output from a bivariate logistic regression with the same two variables. This is output from SPSS LOGISTIC REGRESSION,

```
logistic regression pmrj5 with edf5/method=enter edf5/method=bstep(lr)/print=def/classplot/
          save=pred(lpepmrj5).
```

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	85.359	1	.000
	Block	85.359	1	.000
	Model	85.359	1	.000

Classification Table(a)

			Predicted		
			PMRJ5		Percentage Correct
	Observed		no	yes	
Step 1	PMRJ5	no	136	14	90.7
		yes	37	44	54.3
	Overall Percentage				77.9

a The cut value is .500

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	213.947	.309	.425

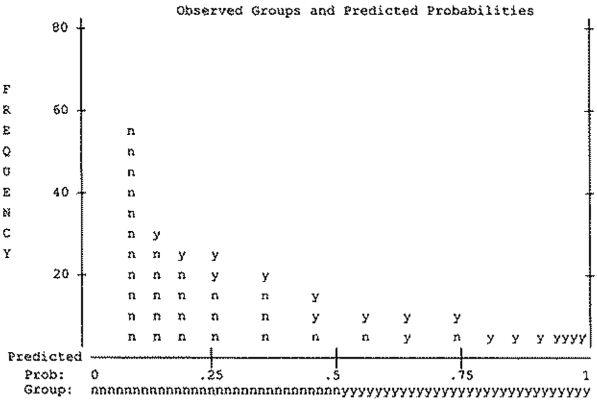
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	EDF5	.407	.058	48.546	1	.000	1.502
	Constant	-5.487	.710	59.732	1	.000	.004

a Variable(s) entered on step 1: EDF5.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.875	6	.130



Predicted Probability is of Membership for yes
The Cut Value is .50
Symbols: n - no
y - yes
Each Symbol Represents 5 Cases.

Figure 1.4. Bivariate Logistic Regression for the Prevalence of Marijuana Use

```
          ++----+----+----+----+----+----+----+----+----+----+----+----+----++
         1+                                      C   C    C     C  $ $ $ $ $ +
          |                                                 L   L            |
          |                                               L                  |
          |                                            L                     |
          |                                          L                       |
      .875+                                                                  +
          |                                        L                         |
          |                                                                  |
          |                                      L                           |
          |                                 C                                |
       .75+                                                                  +
 M        |                                    L                             |
 P        |                                    C                             |
 M        |                                        C   C    C                |
 R        |                                 L                                |
 J    .625+                             C                                    +
 5        |                                                                  |
          |                                                                  |
 M        |                               L                                  |
 L        |                                                                  |
 P      .5+                                                                  +
 E        |                                                                  |
 P        |                             L C                                  |
 M        |                                                                  |
 J        |                                                                  |
 5    .375+                                                                  +
          |                            L                                     |
          |                         C  C                                     |
          |                                                                  |
          |                         L                                        |
       .25+                       C                                          +
          |                                                                  |
          |                       L                                          |
          |                                                                  |
          |                     $                                            |
      .125+                                                                  +
          |                   L                                              |
          |                                                                  |
          |                                                                  |
          |                   C                                              |
         0+                                                                  +
          ++----+----+----+----+----+----+----+----+----+----+----+----+----++
              2.25      6.75     11.25     15.75     20.25     24.75     29.25
            0       4.5         9      13.5        18      22.5        27

                                         EDF5
```

Figure 1.5. Conditional Probabilities Observed (C) and Predicted by Logistic Regression (L)
C: Mean Prevalence of Marijuana Use (MPMRJ5) WITH EDF5 (Exposure to delinquent friends). L: Logistic Regression Prediction of Prevalence of Marijuana Use (MLPEPMRJ5) WITH EDF5. $: Multiple Occurence (Logistic Regression prediction and observed value coincide). 21 cases.

with some deletions but nothing added. The equation for the logit of the prevalence of marijuana use from Figure 1.4 is logit(PMRJ5) = −5.487 + .407(EDF5). There are several other statistics presented in Figure 1.4 that will be discussed in the pages to follow. For the moment, however, note that the presentation of logistic regression results includes (a) some summary statistics for the goodness of fit of the model (Omnibus Tests of Model Coefficients and Model Summary), (b) a comparison of observed and predicted values (or classification) of cases according to whether they do (yes) or do not (no) report using marijuana (Classification Table), (c) the estimated parameters (*B*) of the logistic regression equation, along with other statistics associated with those parameters (Variables in the Equation), and (d) a plot of the observed (yes = 1 or no = 0) and predicted probabilities of "membership" of being marijuana users (Observed Groups and Predicted Probabilities).

Figure 1.5 plots the predicted and observed conditional probabilities (or, equivalently, the conditional means) for the logistic regression equation. The observed conditional probabilities are represented by the letter "C" and the predicted conditional probabilities are represented by the letter "L" for logistic regression. In Figure 1.2, the predicted probabilities from linear regression analysis represented a straight line, and for values of EDF5 greater than 23.5, the predicted conditional probabilities of being a marijuana user were greater than 1. The observed conditional probabilities, unlike the predicted conditional probabilities, leveled off at 1. In Figure 1.5, the conditional probabilities predicted by logistic regression analysis all lie between 0 and 1, and the pattern of the predicted probabilities follows the curve suggested by the observed conditional probabilities, a curve similar to the right half of the curve in Figure 1.3. Just from looking at the pattern, there appears to be a closer correspondence between the observed and predicted conditional means when logistic regression is used to predict the dependent variable.

2. SUMMARY STATISTICS FOR EVALUATING THE LOGISTIC REGRESSION MODEL

When we evaluate a linear regression model, the evaluation typically has three parts. First, how well does the overall model work? Can we be confident that there is a relationship between all of the independent variables, taken together, and the dependent variable, above and beyond what we might expect as a coincidence, attributable to random variation in the sample we analyze? If there is a relationship, how strong is it? Second, if the overall model works well, how important is each of the independent variables? Is the relationship between any of the variables attributable to random sample variation? If not, how much does each independent variable contribute to our ability to predict the dependent variable? Which variables are stronger or weaker, better or worse predictors of the dependent variable? Third and finally, does the form of the model appear to be correct? Do the assumptions of the model appear to be satisfied? In this chapter, we deal with the first question, the overall adequacy of the model. Chapter 3 deals with the contributions of each of the independent variables, and Chapter 4 focuses on testing the assumptions of the model.

In linear regression analysis, we need to know (a) whether knowing the values of all of the independent variables put together allows us to predict the dependent variable any better than if we had no information on any of the independent variables and, if so, (b) how well the independent variables as a group explain the dependent variable. For logistic regression, we also may be interested in the frequency of correct as opposed to incorrect predictions of the exact value of the dependent variable, in addition to how well the model minimizes errors of prediction. In linear regression, when the dependent variable is assumed to be measured on an interval or ratio scale, it would be neither alarming nor unusual to find that none of the predicted values of the dependent variable exactly matched the observed value of the dependent variable. In logistic regression, with a finite number (usually only two) of possible values of the dependent variable, we may sometimes be more concerned with whether the predictions are correct or incorrect than with how close the predicted values (the predicted conditional means, which are equal to the predicted conditional probabilities) are to the observed (0 or 1) values of the dependent variable.

2.1. R^2, F, and Sums of Squared Errors

In linear regression analysis, evaluation of the overall model is based on two sums of squares. If we were concerned with minimizing the sum of the squared errors of prediction and if we knew only the values of the dependent variable (but not the cases to which those values belonged), we could minimize the sum of the squared errors of prediction by using $\overline{Y}$, the mean of Y, as the predicted value of Y for all cases. The sum of squared errors based on this prediction would be $\sum(Y_j - \overline{Y})^2$, the *Total Sum of Squares* (SST). If the independent variables are useful in predicting Y, then $\widehat{Y}_j$, the value of Y predicted by the regression equation (the conditional mean of Y) will be a better predictor than $\overline{Y}$ of the values of Y, and the sum of squared errors $\sum(Y_j - \widehat{Y}_j)^2$ will be smaller than the sum of squared errors $\sum(Y_j - \overline{Y})^2$. $\sum(Y_j - \widehat{Y}_j)^2$ is called the *Error Sum of Squares* (SSE) and is the quantity OLS selects parameters $(\beta_1, \beta_2, \ldots, \beta_k)$ to minimize. A third sum of squares, the *Regression Sum of Squares* (SSR) is simply the difference between SST and SSE: SSR = SST − SSE.

It is possible in a sample of cases to get an apparent reduction in error of prediction by using the regression equation instead of $\overline{Y}$ to predict the values of Y_j, even when the independent variables are really unrelated to Y. This occurs as a result of sampling variation, that is, random fluctuations in sample values that may make it appear as though a relationship exists between two variables when there really is no relationship. The multivariate F ratio is used to test whether the improvement in prediction using $\widehat{Y}$ instead of $\overline{Y}$ is attributable to random sampling variation. Specifically, the multivariate F ratio tests two equivalent hypotheses: $H_0 : R^2 = 0$ and $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. For OLS linear regression, the F ratio with N cases and k independent variables can be calculated as

$$F = [\mathrm{SSR}/k]/[\mathrm{SSE}/(N - k - 1)] = (N - k - 1)\mathrm{SSR}/(k)\mathrm{SSE}.$$

The *attained statistical significance* (p) associated with the F ratio indicates the probability of obtaining an R^2 as large as the observed R^2, or β coefficients as large as the observed β coefficients, *if the null hypothesis is true*. If p is small (usually less than .05, but other values of p may be chosen), then we reject the null hypothesis and conclude that there is a relationship between the independent variables and the dependent variable that cannot be attributed to chance. If p is large, then we "fail to reject the null hypothesis" and conclude that there is insufficient evidence to be sure that the variance explained by the model is not attributable to random sample variation. This does not mean that we conclude that there is no relationship, only that if there is a relationship, we have insufficient evidence to be confident that it exists.

The coefficient of determination, or R^2, or "explained variance" (really, the proportion of the variance that is explained) is an indicator of *substantive* significance, that is, whether the relationship is "big enough" or "strong enough" for us to be concerned about it. R^2 is a *proportional reduction in error* statistic. It measures the proportion (or, multiplied by 100, the percentage) by which use of the regression equation reduces the error of prediction relative to predicting the mean, $\overline{Y}$. R^2 ranges from 0 (the independent variables are no help at all) to 1 (the independent variables allow us to predict the individual values Y_j perfectly). R^2 is calculated as $R^2 = \text{SSR/SST} = (\text{SST} - \text{SSE})/\text{SST} = 1-(\text{SSE/SST})$. The F ratio and R^2 also can be expressed as functions of one another: $F = [R^2/k]/[(1 - R^2)/(N - k - 1)]$ and $R^2 = kF/(kF + N - k - 1)$.

It is possible for a relationship to be statistically significant ($p \leq .0001$) but for R^2 not to be substantively significant (for example, $R^2 \leq .005$) for a large sample. If the independent variables explain less than one-half of 1% of the variance in the dependent variable, we are unlikely to be very concerned with them, even if we are relatively confident that the explained variance cannot be attributed to random sample variation. It is also possible for a relationship to be substantively significant (for example, $R^2 \geq .4$), but not statistically significant for a small sample. Even though the relationship appears to be moderately strong (an explained variance of .40 or, equivalently, a 40% reduction in errors of prediction), there may not be enough cases for us to be confident that this result cannot be attributed to random sampling variation.

2.2. Goodness of Fit: G_{M}, R^2_{L}, and the Log Likelihood

Close parallels to F and R^2 exist for the logistic regression model. Just as the sum of squared errors is the criterion for selecting parameters in the linear regression model, the *log likelihood* is the criterion for selecting parameters in the logistic regression model. In presenting information on the log likelihood, however, statistical packages usually present not the log likelihood itself, but the log likelihood multiplied by −2, for reasons noted subsequently. For convenience, the log likelihood multiplied by −2 will be abbreviated as −2LL. Whereas the log likelihood is negative, −2LL is positive, and larger values indicate worse prediction of the dependent variable. The value of −2LL for the logistic regression model with only the intercept included can be calculated in SPSS LOGISTIC REGRESSION by adding the chi-square for the model in the Omnibus Tests of Model Coefficients table plus the −2 log likelihood in the Model Summary table (see Figure 1.4). In SPSS NOMREG and PLUM, to be discussed in more detail later, it is the −2 log likelihood for intercept only in the Model Fitting Information table. In SAS, it is designated as −2 LOG L in the column "Intercept Only" in the output from SAS PROC LOGISTIC. The intercept-only or initial −2LL, hereafter designated D_0 to indicate that is the −2 log-likelihood statistic with none (zero) of the independent variables in the equation, is analogous to the total sum of squares (SST) in linear regression analysis. For a dichotomous dependent variable (coded as 0 or 1), if $n_{Y=1}$ is the number of cases for which $Y = 1, N$ is the total number of cases, and $\mathrm{P}(Y = 1) = n_{Y=1}/N$ is the probability that Y is equal to 1, then

$$\begin{aligned} D_0 &= -2\{n_{Y=1}\ln[\mathrm{P}(Y = 1)] + (N - n_{Y=1})\ln[1 - \mathrm{P}(Y = 1)]\} \\ &= -2\{(n_{Y=1})\ln[\mathrm{P}(Y = 1)] + (n_{Y=0})\ln[\mathrm{P}(Y = 0)]\}. \end{aligned}$$

The value of −2LL for the logistic regression model that includes the independent variables as well as the intercept is designated as −2 log likelihood in the Model Summary table in the output for SPSS LOGISTIC REGRESSION, as −2 log likelihood for the final model in the Model Fitting Information table in SPSS NOMREG and PLUM, and as −2 LOG L in the "Intercept and Covariates" column in SAS PROC LOGISTIC. Hereafter, this −2LL statistic will be referred to as D_M for the full model. D_M is analogous to the error sum of squares (SSE) in linear regression analysis. The most direct analogue in logistic regression analysis to the regression sum of squares (SSR in linear regression) is the difference between D_0 and D_M, that is, $(D_0 - D_\mathrm{M})$. This difference is called the Model chi-square (in the Omnibus Tests table) in SPSS LOGISTIC REGRESSION, or the chi-square for the final model (in the Model Fitting Information table) in SPSS NOMREG and PLUM, or −2 LOG L in the column "Chi-Square for Covariates" in SAS PROC LOGISTIC. Hereafter, it will be referred to as G_M, or the model χ^2.

In logistic regression (and in other general linear models), the difference between two log likelihoods, when multiplied by −2, can be interpreted as a χ^2 statistic if they come from two different models, one of which is *nested* within the other (McCullagh & Nelder, 1989). One model is nested within another if the first model contains some, but not all, of the predictors in the second model and contains no predictors that are not included in the second model. In other words, the predictors in the first model are a proper subset of the predictors in the second. G_{M} can be straightforwardly interpreted as the difference between a first model that contains only an intercept and a second model that contains the intercept plus one or more variables as predictors. Treated as a chi-square statistic, G_{M} provides a test of the null hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ for the logistic regression model. If G_{M} is statistically significant ($p \leq .05$), then we reject the null hypothesis and conclude that information about the independent variables allows us to make better predictions of $\mathrm{P}(Y = h)$ (where h is some specific value, usually 1, usually for a dichotomous dependent variable) than we could make without the independent variables. G_{M} is thus analogous to the multivariate F test for linear regression as well as the regression sum of squares.

Designated the "deviance" by McCullagh and Nelder (1989) and others (a term with, at best, mixed meanings when the substantive example is marijuana use and that I will avoid to the extent possible hereafter), D_{M} has historically been used as a measure of "goodness of fit," which is essentially a test for the statistical significance of the variation *unexplained* by the logistic regression model and is akin to testing for the statistical significance of unexplained variance in an OLS regression model. If a cliche will help, G_{M} asks how full the cup is (how much improvement the predictors make in predicting the dependent variable), while D_{M} asks how empty the cup is (how much improvement is needed before the predictors provide the best possible prediction of the dependent variable). While G_{M} compares the intercept-only model with the full model (the model that includes all the predictors), D_{M} compares the full model with a *saturated* model (a model that includes all predictors plus all possible interactions among them). In previous versions of SPSS LOGISTIC REGRESSION (and in the first edition of this monograph), D_{M} was assumed to have an approximately χ^2 distribution and was assigned a level of statistical significance. The problem with using D_{M} as a χ^2 statistic lies in the fact that there are different ways to define a saturated model, resulting in different values for D_{M} and different degrees of freedom (Simonoff, 1998).

Briefly (and bypassing some detail), as explained by Simonoff (1998), one approach (the one taken in SPSS LOGISTIC REGRESSION and SAS PROC LOGISTIC) is to consider each case as independent (casewise approach), and contributing 1 degree of freedom. The alternative is to consider each combination of values of the predictors, or each *covariate pattern*, as a separate cell in a crosstabulation (contingency table approach), and to calculate degrees of freedom based on the number of covariate patterns (cells in the table) rather than the number of individuals. In either approach, if the number of cases per covariate pattern (cell) is too small or if there are many empty cells, D_{M} will not generally have a χ^2 distribution and it would be inappropriate to use it as a χ^2 statistic to test goodness of fit (McCullagh & Nelder, 1989; Simonoff, 1998). If there is a large number of cases relative to the number of covariate patterns and sufficient cases per covariate pattern, it is possible to define an appropriate saturated model and to calculate a deviance statistic that will have a χ^2 distribution and the correct degrees of freedom based on the contingency table approach. This is done in SPSS NOMREG and PLUM, both of which can be used to analyze dichotomous as well as nominal or ordinal variables with more than two categories. In NOMREG and PLUM, the Goodness-of-Fit table provides Pearson and deviance χ^2 statistics, the latter based on the −2 log likelihood.

For casewise data, it is still possible to construct a goodness-of-fit index. One commonly available index for dichotomous dependent variables is Hosmer and Lemeshow's (1989) goodness-of-fit index $\widehat{C}$, which can be included in the output for SPSS LOGISTIC REGRESSION or SAS PROC LOGISTIC. Hosmer and Lemeshow's goodness-of-fit index was designed primarily as an alternative to avoid the problems associated with using D_{M} as a goodness-of-fit index for casewise data, and it proceeds by collapsing the data into deciles based on the probability of having the characteristic of interest (for example, being a marijuana user). Other possible goodness-of-fit indices include the score statistic, the Akaike information criterion (AIC), and the Schwartz criterion (a modification of the AIC), all of which are provided in SAS PROC LOGISTIC. The score statistic is, like G_{M}, a test of the statistical significance of the combined effects of the independent variables in the model. The AIC and the Schwartz criterion, which are briefly discussed in Bollen (1989), are two related indices used to compare models, rather than to provide absolute tests of adequacy of fit. It is possible to compare the AIC or the Schwartz criterion for the fitted model with the AIC or the Schwartz criterion for the model with only the intercept, but this provides little more information than G_M.

For some researchers, particularly those who have a strong background in log-linear models or general linear models, or a perspective that is more theoretical than applied, goodness of fit will be an important consideration. Given the goal of the logistic regression model (prediction of a single dependent variable), and consistent both with an applied focus and with the analogy between linear and logistic regression, it seems advisable for most purposes to focus here primarily on G_M.

2.2.1. Measures of Multiple Association Between the Independent Variables and the Dependent Variable

Several analogues to the linear regression R^2 have been proposed for logistic regression. For general reviews, see Hagle and Mitchell (1992), Menard (2000), and Veall and Zimmerman (1996). Here the focus is on R^2 analogues that are commonly used in general purpose statistical packages such as SAS and SPSS, and on some general categories of coefficients of determination with which they may be compared. If we maintain the analogy between the −2LL statistics for logistic regression and the sums of squares for linear regression analysis, the most natural choice, directly analogous to SSR/SST, is the likelihood ratio R^2, $R^2_L = G_M/(D_0) = G_M/(G_M + D_M)$ (McFadden, 1974; see also Agresti, 1990, pp. 110–111; DeMaris, 1992, p. 53; Hosmer & Lemeshow, 1989, p. 148; Knoke & Burke, 1980, p. 41; Menard, 2000). R^2_L is a *proportional reduction in −2LL* or a *proportional reduction in the absolute value of the log-likelihood* measure, where the −2LL or the absolute value of the log likelihood—the quantity being minimized to select the model parameters—is taken as a measure of "variation" (Nagelkerke, 1991), not identical but analogous to the variance in OLS regression. R^2_L indicates how much inclusion of the independent variables in the model reduces the variation, as measured by D_0. The variation is between 0 (for a model in which $G_M = 0$, $D_M = D_0$, and the independent variables are useless in predicting the dependent variable) and 1 (for a model in which $G_M = D_0$, $D_M = 0$, and the model predicts the dependent variable with perfect accuracy). R^2_L can be obtained directly from the output for SPSS

NOMREG and PLUM, where it is presented as the McFadden R^2 in the Pseudo-R^2 table. Curiously it is not included (as of this writing) in SPSS LOGISTIC REGRESSION. Instead, in SPSS LOGISTIC REGRESSION and SAS PROC LOGISTIC, it must be computed by hand from the information provided (as described previously) on D_0 (or D_{M}) and G_{M}.[7]

Two measures used in the current versions of SPSS and SAS are (1) the geometric mean squared improvement per observation $R^2_{\mathrm{M}} = 1 - (L_0/L_{\mathrm{M}})^{2/N}$, where L_0 is the likelihood function for the model that contains only the intercept, L_{M} is the likelihood function that contains all the predictors, and N is the total number of cases (Cox & Snell, 1989; Maddala, 1983, pp. 39–40), and (2) an adjusted geometric mean squared improvement per observation R^2_N (Cragg & Uhler, 1970; Maddala, 1983, p. 40; Nagelkerke, 1991). The unadjusted measure cannot have a value of 1, even for a model that fits the data perfectly. The adjusted measure permits a value of 1 by dividing by the maximum possible value of R^2_{M} for a particular dependent variable in a particular data set: $R^2_{\mathrm{N}} = [1 - (L_0/L_{\mathrm{M}})^{2/N}]/[1 - (L_0)^{2/N}] = R_{\mathrm{M}}{}^2$/(maximum possible R^2_M). In SPSS LOGISTIC REGRESSION, R^2_{M} and R^2_{N} are presented, respectively, as the Cox–Snell and Nagelkerke R^2 measures in the Model Summary table or as the Cox–Snell and Nagelkerke pseudo-R^2 measures in the Pseudo-R^2 table in SPSS NOMREG and PLUM. In SAS PROC LOGISTIC, they are simply referred to as the R^2 and adjusted R^2.

A family of alternatives to R^2_{L} includes the pseudo-R^2 or contingency coefficient R^2_{C}, which was proposed by Aldrich and Nelson (1984) in their discussion of logit and probit models, the Wald R^2_{W} (Magee, 1990), and the McKelvey and Zavoina (1975) R^2_{MZ}. In the notation used in this monograph, if N is the number of cases, $R^2_{\mathrm{C}} = G_{\mathrm{M}}/(G_{\mathrm{M}}+N)$. Similarly, the Wald $R^2_{\mathrm{W}} = W/(W+N)$, where W is the multivariate Wald statistic. The McKelvey–Zavoina $R^2_{\mathrm{MZ}} = s^2_{\hat{Y}}/(s^2_{\hat{Y}}+1)$ for the probit model (the context in which it was originally developed) or $R^2_{\mathrm{MZ}} = s^2_{\hat{Y}}/(s^2_{\hat{Y}} + \pi^2/3)$ for a logit or logistic regression model, where $s^2_{\hat{Y}}$ is the variance in $\widehat{Y}$ (the predicted value of Y), and 1 and $\pi^2/3$ are the standard deviations for the standard normal and logistic distributions, respectively. These measures share the common feature that they cannot attain a value of 1, even for a perfect model fit. Hagle and Mitchell (1992) suggested a correction for Aldrich and Nelson's pseudo-R^2 that allows it to vary from 0 to 1; in principle, this approach could also be applied to the Wald and McKelvey–Zavoina measures.

Hagle and Mitchell also noted that the corrected R_C^2 provided a good approximation for the OLS regression R^2, and Veall and Zimmerman noted the same with respect to the McKelvey–Zavoina R_{MZ}^2, *when the dichotomous dependent variable represents a latent interval scale*. In this instance, however, there are several other alternatives, including the possibility of using a linear probability model (because the restriction of values to a dichotomy is really artificial for a latent interval scale), using polychoric correlation and weighted least-squares estimation in the context of a more complex structural equation model (Jöreskog & Sörbom, 1993), and using R^2 itself to measure the strength of the association between the observed and predicted values of the dependent variable.

The use of R^2, the familiar coefficient of determination from OLS linear regression analysis, has received relatively little attention in the literature on logistic regression analysis. (For an exception, see Agresti, 1990, pp. 111–112.) Its utility in logistic regression has been questioned because, unlike R_L^2 and Aldrich and Nelson's pseudo-R^2, it is not based on the criteria used to select the model parameters. Also, if the dichotomous dependent variable is assumed to be an indicator for an unmeasured latent variable, R^2 provides a biased estimate of the explained variance. There are certain advantages to the use of R^2, not instead of R_L^2, but as a supplemental measure of association between the independent variables and the dependent variable. First, using R^2 permits direct comparison of logistic regression models with linear probability, analysis of variance, and discriminant analysis models when predicting the observed value (instead of predicting the observed probability that the dependent variable is equal to that value) is of interest. Second, R^2 is useful in calculating standardized logistic regression coefficients, a topic to be covered in the next chapter. Third, R^2 is relatively easy to calculate using existing statistical software.

To calculate R^2 for logistic regression, assume that the dependent variable is Y and that you want to name the variable that represents the value of Y predicted by the logistic regression model LPREDY. In SPSS and SAS, to obtain R^2, it is necessary to save the predicted values of the dependent variable from SPSS LOGISTIC REGRESSION [using SAVE = PRED(LPREDY)] or from SAS PROC LOGISTIC [using OUTPUT PRED = LPREDY]. Next, use a bivariate or multiple regression routine (such as SPSS REGRESSION or SAS PROC REG) to calculate R^2. Alternatively, use any analysis of variance routine that calculates η^2 or η (SPSS MEANS or ANOVA; SAS PROC GLM or ANOVA) with the *observed* value of the dependent variable, Y, as the *independent* variable and the *predicted* value of the dependent variable, LPREDY, as the *dependent* variable. Because there are only two variables (the observed values of Y as one variable, the predicted values of Y as the other), $\eta^2 = R^2$ and the two variables may be used interchangeably. Although for η^2 this role switching between the dependent variable and its predicted value (which is based on the values of the independent variables) may seem strange for η^2, it exactly parallels the method for calculating canonical correlation coefficients in discriminant analysis (Klecka, 1980).

Based on research on the properties of the different proposed measures, I have suggested (Menard, 2000) that R_{L}^2 is the most appropriate for logistic regression, based on several considerations.[8] First and most importantly, R_{L}^2 is conceptually closest to the OLS R^2 insofar as it reflects a proportional reduction in the quantity actually being minimized (−2LL; equivalently, the log likelihood is being maximized), in contrast to R^2, R_{W}^2, and R_{MZ}^2. Also, unlike measures that depend on the sample size as well as the log likelihood or −2LL (R_{M}^2, R_{N}^2, R_{C}^2), R_{L}^2 depends *only* on the quantity being maximized or minimized. Second, R_{L}^2 is not sensitive to the *base rate*, the proportion of cases that have the attribute (for example, being or not being a marijuana user) being studied. Evidence indicates that R_{M}^2, R_{N}^2, R_{C}^2, and R^2 all have the undesirable property that their value increases as the base rate (whichever is smaller, $n_{Y=1}/N$ or $n_{Y=0}/N$) increases from 0 to .50, absurdly suggesting that one could, in effect, substitute the sample size for one of these coefficients of determination as a measure of explained variation (Menard, 2000, p. 23). Third, R_{L}^2, unlike the unadjusted versions of R_{W}^2, R_{C}^2, and R_{MZ}^2, varies between 0 and 1, where 0 represents no predictive utility for the independent variables and 1 represents perfect prediction. Fourth, as noted by Veall and Zimmerman (1996), R_{L}^2 works as well for polytomous nominal or ordinal dependent variables as for dichotomous dependent variables, in contrast to the variance-based measures R_{MZ}^2 and R^2.

2.3. Predictive Efficiency: λ_{p}, τ_{p}, ϕ_{p}, and the Binomial Test

In addition to statistics regarding goodness of fit, logistic regression programs commonly print classification tables that indicate the predicted and observed values of the dependent variable for the cases in the analysis. These tables resemble the contingency tables produced by SPSS CROSSTABS and SAS PROC FREQ. In most instances, we will be more interested in how well the model predicts probabilities, $P(Y_j = 1)$. In other cases, however, we may be more interested in the accurate prediction of group membership, so the classification tables may be of as much or more interest than the overall fit of the model. There is no consensus at present on how to measure the association between the observed and predicted classification of cases based on logistic regression or related methods such as discriminant analysis. There are, however, several good suggestions that can easily be implemented to provide summary measures for classification tables. The best options for analyzing the prediction tables provided by logistic regression packages involve *proportional change in error* measures of the form

$$\text{predictive efficiency} = \frac{(\text{errors without model}) - (\text{errors with model})}{(\text{errors without model})}, \qquad [2.1]$$

which is a *proportional change in error* formula. If the model improves our prediction of the dependent variable, this formula is the same as a *proportional reduction in error* (PRE) formula. It is possible under some circumstances, however, that a model actually will do worse than chance at predicting the values of the dependent variable. When that occurs, the predictive efficiency is negative and we have a proportional *increase* in error. The errors with the model are simply the number of cases for which the predicted value of the dependent variable is incorrect. The errors without the model differ for the three indices and depend on whether we are using a prediction, classification, or selection model.

2.3.1. Prediction, Classification, and Selection Models

In prediction models, the attempt is made to classify cases according to whether they satisfy some criterion, such as success in college, absence of behavioral or emotional problems in the military, or involvement in illegal behavior after release from prison. In prediction models, there are no *a priori* constraints on the number or proportion of cases predicted to have or not have the specified behavior or characteristic. In principle, it is possible (but not necessary) to have the same number of cases *predicted* to be "positive" (having the behavior or characteristic, e.g., "successes") and "negative" (not having the behavior or characteristics, e.g., "failures") as are *observed* to be positive and negative. That is, there is nothing that constrains the *marginal distributions* (the number or proportion of cases in each category, positive or negative) of predicted and observed frequencies to be equal or unequal. In particular, all cases may be predicted to belong to the same category, that is, the sample or population may be *homogeneous*. In practical terms, prediction models are appropriate when identical treatment of all groups ("lock 'em all up" or "let 'em all go") is a viable option.

In classification models, the goal is similar to that of prediction models, but there is the added assumption that the cases are truly heterogeneous. Correspondingly, the evaluation of a classification model imposes the constraint that the model should classify as many cases into each category as are actually observed in each category. The proportion or number of cases observed to be in each category (the *base rate*) should be the same as the proportion or number of cases predicted to be in each category. To the extent that a model fails to meet this criterion, it fails as a classification model. Complete homogeneity is an unacceptable solution for a classification model. Practically speaking, classification models are appropriate when heterogeneity is assumed, and identical treatment of all groups is not a viable option.

In selection models (Wiggins, 1973), the concern is with "accepting" or "rejecting" cases for inclusion in a group, based both on whether they will satisfy some criterion for success in the group and on the minimum required, maximum allowable, or specified number of cases that may (or must) be included in the group. In selection models, the proportion of cases observed to be successful (the *base rate* again) may or may not be equal to the proportion of cases accepted or selected for inclusion in the group (the *selection ratio*). For example, a company may need to fill 20 positions from a pool of 200 applicants. The selection ratio will be 20/200 = .10 (10%) regardless of whether the base rate (the observed probability of success on the job) is 5% or 20%, half or twice the selection ratio. The classification tables provided in logistic regression packages may naturally be regarded as prediction or classification models. They may be used to construct selection models, but they must be altered (unless, purely by coincidence, the selection ratio turns out to be equal to the base rate) so that the correct number of cases is selected.

2.3.2. Common Measures of Association for Contingency Tables as Indices of Predictive Efficiency

Among the various measures that have been considered as indices of predictive efficiency are several measures of association that are commonly employed to analyze contingency tables: ϕ, Goodman and Kruskal's γ, κ, the contingency coefficient, Pearson's r, and the odds ratio (Farrington & Loeber, 1989; Mieczkowski, 1990; Ohlin & Duncan, 1949). The problem with using common contingency table measures of association to analyze 2×2 or larger prediction tables lies in the distinction between (1) the strength of a relationship between an independent variable X and a dependent variable Y, and (2) the strength of the relationship between predicted group membership $E(Y_j)$ and observed group membership Y_j. These differences are illustrated in Figure 2.1. Table A in Figure 2.1 represents the general format to be used throughout this section in designating cell and marginal frequencies in 2×2 tables, Table B represents the hypothetical relationship between ethnicity and political orientation, and Table C illustrates the hypothetical relationship between predicted and observed political orientation.

Although Tables B and C are numerically identical, the inferences to be drawn from them are very different. In Table B, knowledge of ethnicity allows us to predict political orientation with a *proportional reduction in error* (PRE) (Bohrnstedt & Knoke, 1994, p. 164; Costner, 1965) of .20 according to Goodman and Kruskal's λ or .04 according to Goodman and Kruskal's τ. In Table C, the PRE is the same, but only if we predict the *opposite* of what the hypothetical model predicts. Actually, the model does *worse* than chance in predicting political orientation (a situation that may arise naturally with skewed data or with the application of a prediction model developed from one set of data to another set of data). If every case were *misclassified*, both λ and τ would have a value of 1.00 for Table C; they would make no distinction between perfectly accurate classification and perfect *misclassification*. Pearson's r and its equivalents for 2×2 tables, Kendall's τ and ϕ, when it is calculated as

$$\phi = (ad - bc)/\sqrt{(a+b)(a+c)(b+d)(c+d)}$$

would indicate misclassification with a negative sign and may be interpreted as PRE measures when squared. For larger tables with

Table A: Standard Format for Prediction Tables

		Predicted Y		
		Positive (success)	Negative (failure)	
Observed Y	Positive (success)	a	b	$a + b$
	Negative (failure)	c	d	$c + d$
		$a + c$	$b + d$	$a + b + c + d$

Table B: Ethnicity and Political Orientation

		X: Ethnicity		
		European	Non-European	
Political Orientation	Conservative	20	30	50
	Liberal	30	20	50
		50	50	100

Table C: Predicted and Observed Political Orientation

		Predicted Political Orientation		
		Conservative	Liberal	
Observed Political Orientation	Conservative	20	30	50
	Liberal	30	20	50
		50	50	100

Figure 2.1. Association Versus Prediction. For Tables B and C, Goodman and Kruskal's $\lambda = 0.20$; Goodman and Kruskal's $\tau = \phi^2 = r^2 = 0.04$

unordered categories, however, Pearson's r and Kendall's τ cannot be used, and ϕ becomes Cramer's V, which no longer has a PRE interpretation. The odds ratio may also be used for 2 × 2 tables, but for larger tables, two or more odds ratios must be calculated, and the odds ratio no longer provides a single summary measure of accuracy of prediction. On the whole, it does not appear that the application of common measures of association for contingency tables to predictive tables provides a straightforward or general solution to the problem of estimating accuracy of prediction. Pearson's r and r^2, or ϕ and ϕ^2, are reasonable indices for use with 2 × 2, but not larger, tables, as long as we remember to interpret them contingent upon the sign of r or ϕ.

2.3.3. λ_p, τ_p, *and* ϕ_p

Equation 2.1 provides a basic form for indices of predictive efficiency. Errors with the model are simply the number of cases misclassified when we use the model and are analogous to the error sum of squares. Errors without the model are analogous to the total sum of squares and depend on whether we are using a prediction model, a classification model, or a selection model. For a prediction model, the approach most closely analogous to linear regression (with an interval level dependent variable) is to use the mode of the dependent variable as the predicted value for all cases (analogous to using the mean in linear regression). This method of defining errors without the model is the same as the one used in defining Goodman and Kruskal's λ for contingency tables with nominal variables, and gives us an index first proposed by Ohlin and Duncan (1949). Because of the similarity to Goodman and Kruskal's λ, the index is here referred to as λ_p (lambda-p) where the subscript p refers to its use with prediction tables.

Lambda-p is a PRE measure like R^2 when it is positive, but if the model does worse than predicting the mode, λ_p may be negative, indicating the proportional *increase* in error. The possible values of λ_p vary depending on the marginal distributions. In general, the full range of possible values for λ_p in all tables with N cases is from $1-N$ to 1.

For a classification model, an appropriate definition of the expected error without the model is

$$\text{errors without model} = \sum_{i=1}^{N} f_i[(N - f_i)/N],$$

where N is the sample size and f_i is the number of cases observed in category i. This is the same formula for error without the model as is used for Goodman and Kruskal's τ. An index based on this definition of errors without the model was proposed by Klecka (1980) for use with discriminant analysis models. Parallel to λ_p, Klecka's index will be referred to as τ_p (tau-p) or τ for prediction tables.

Like λ_p, τ_p is a measure of change. Unlike λ_p, τ_p requires that even in the estimation of error without the model, cases must be separated into distinct groups or categories, and not all placed in the same category. In effect, τ_p adjusts the expected number of errors for the base rates of classification. Accuracy of prediction is thus secondary, subject to the a priori assumption of heterogeneity. As with λ_p, a value of 1 for τ_p indicates that all cases are correctly classified; a negative value for τ_p indicates that the prediction model does worse than expected (based on the observed marginal distribution) in predicting the classification of cases. Secondary properties of τ_p include the fact that $\tau_p \geq \lambda_p$ for a dichotomous dependent variable because the number of errors without the model for τ_p will be equal to or larger than the number of errors without the model for λ_p. For tables with equal marginal distributions, τ_p varies between -1 and $+1$, but the maximum value of τ_p is less than 1 when the marginal distributions are unequal. In the worst possible case, with extremely skewed and inconsistent marginal distributions, the minimum value of τ_p is equal to $1 - N^2/2(N-1)$. The range of τ_p is not constant, but varies from 1 to 2 for different marginal distributions. A smaller range occurs for models in which τ_p is negative and large in absolute value.

It is also possible to construct a proportional change in error measure of accuracy of prediction for selection models. For such a measure, the error with the model will be $b + c$, just as it is for λ_p and τ_p. Error with the model should depend on both the base rate, $B = (a+b)/N$, and the selection ratio, $S = (a+c)/N$. Given B, S, and N, we know the expected value of cell a (Table A, Figure 2.1): $E(a) = BSN$. Because a 2×2 prediction table has only 1 degree of freedom, once the expected value of a is known, then given the marginal distribution, the expected value of all of the other cells is known and is identical to the expected values used to calculate the χ^2 statistic. The expected error is $E(b+c) = [(a+b)(b+d)/N + (c+d)(a+c)/N]$. Plugging these values into the PRE formula, we obtain a proportional change in error measure[9]

$$\begin{aligned}\phi_p &= \frac{(a+b)(b+d)/N + (c+d)(a+c)/N - (b+c)}{(a+b)(b+d)/N + (c+d)(a+c)/N} \\ &= \frac{(a+b)(b+d) + (c+d)(a+c) - N(b+c)}{(a+b)(b+d) + (c+d)(a+c)} \\ &= \frac{ad - bc}{.5[(a+b)(b+d) + (c+d)(a+c)]}.\end{aligned}$$

For tables with equal marginal distributions, ϕ_p has a maximum value of +1. In general, it varies between −1 and +1, but the actual maximum, minimum, and range depend on the marginal distributions. As long as errors without the model are calculated as the sum of the expected frequencies in cells *b* and *c* (Table A, Figure 2.1), and errors with the model are calculated as the sum of the observed frequencies in cells *b* and *c*, ϕ_p can be extended to tables larger than 2×2 and still retain a proportional change in error interpretation. For 2×2 tables, it can be shown that $|\phi_p| \leq |\phi|$ and that ϕ_p has the same sign as ϕ and Pearson's *r* (the numerator is the same, $ad - bc$, as for ϕ). When all cases are correctly predicted, $\phi_p = 1$; otherwise, $\phi_p < 1$, even when the maximum possible number of cases for a given set of marginals is correctly classified.[10]

2.3.4. *Statistical Significance of* λ_p, τ_p, *and* ϕ_p

Lambda-p, tau-p, and phi-p are analogous to R^2 as measures of substantive significance. For statistical significance, an analogue to the *F* test is the normal approximation to the binomial test. Let N = total sample size, P_e = (errors without model)/N, and p_e = (errors with model)/N. The binomial statistic *d* may then be computed as

$$d = (P_e - p_e)/\sqrt{P_e(1 - P_e)/N}$$

and *d* is approximately normally distributed (Bulmer, 1979).[11] Note that what is being compared is not the proportion of cases in each category, but the *proportion of cases correctly or incorrectly classified* by the model. This test is the same for λ_p, τ_p, and ϕ_p, for predictive, classification, and selection models. Only the definition of errors without the model differs.

In the proposed test of statistical significance, the value of the observed classification is taken as given; the test indicates whether the proportion predicted incorrectly with the model (which is, by assumption, dependent on the model, and thus variable) differs significantly from the proportion incorrectly predicted without the model (which is dependent only on the marginal distribution, not on the model, and thus assumed to be fixed). This form of the binomial test, which explicitly uses the expected number of errors as the criterion by which the number of errors generated by the model is to be judged, is preferable to the binomial test for a *difference* of two proportions (Bulmer, 1979, p. 145), which assumes that the two proportions (errors with the model and errors without the model) are based on separate samples (possibly of unequal size), a condition that is clearly not met when comparing observed and predicted classifications taken from the classification tables generated by logistic regression, or expected and actual errors, both of which are derived from these tables. The binomial test for the difference of two proportions may, however, be useful if we want to test whether the overall predictive accuracy (percent correctly predicted) is statistically significantly different for two separate prediction models. Even in this situation, however, we would want a separate test to indicate whether either or both of the prediction models was significantly better than chance in reproducing the observed classification of cases.

2.3.5. Other Proposed Indices of Predictive Efficiency

Maddala (1983, pp. 76–77) reviewed three indices of predictive efficiency proposed in econometric literature. One he dismissed (appropriately) as being unable to distinguish a perfectly accurate model from a perfectly inaccurate model, a characteristic it shares with many of the measures of association commonly used in the analysis of contingency tables. A second index, which considered both "first best" and "second best" guesses, was more akin to indices of goodness of fit insofar as a "near miss" (the second most likely category, according to the prediction model) was credited as an accurate choice. The third index typically varied from -1 to $+.50$, depending on the marginal distribution, and produced values similar to ϕ and Pearson's r. Because it lacks a PRE interpretation, it appears to have little or no advantage over ϕ or Pearson's r.

Loeber and Dishion (1983), Copas and Loeber (1990), and Farrington and Loeber (1989) proposed a measure they called relative improvement over chance (RIOC). Although Loeber and his colleagues applied the RIOC to the analysis of prediction and classification tables, the measure corrects for differences between the base rate and the selection ratio, an approach appropriate for selection models rather than classification or prediction models. This measure is identical to the coefficient ϕ', the ϕ coefficient corrected for the marginal distribution. Unlike ϕ, ϕ' has no PRE interpretation. The measure varies between -1 (for perfectly inaccurate prediction, *if cells b and* *c in Table A of Figure 2.1 are both nonzero*) and +1. If *either one* of the cells (*b* or *c*) that contains incorrect predictions is equal to 0, RIOC = 1, *regardless of how small a proportion of the cases are correctly classified* (a problem it shares with Yule's *Q*, a measure of association sometimes used for contingency tables). Even if over 90% of the cases are misclassified, RIOC may have a value of 1.[12]

2.3.6. Comparing Indices of Predictive Efficiency

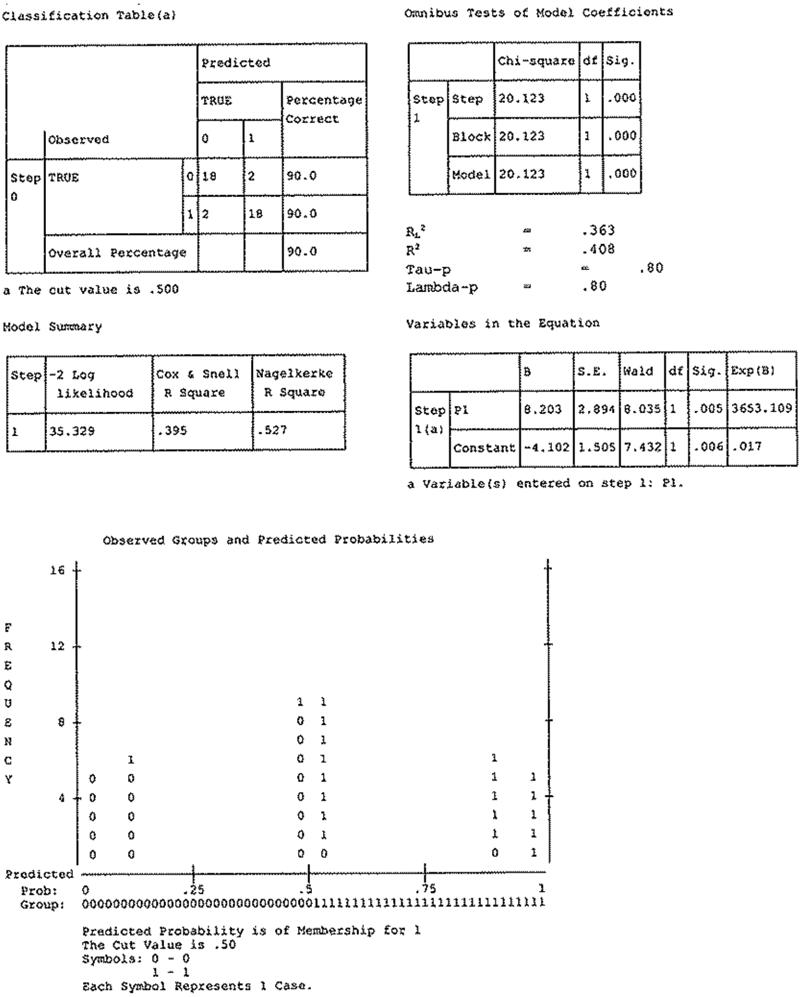
Menard (2000) provided empirical evidence that when the base rate alone is manipulated, λ_{p} and ϕ_{p} are highly correlated with the base rate, but τ_{p} is not. Soderstrom and Leitner (1997) compared λ_{p}, τ_{p}, and ϕ_{p}, and also the RIOC and percentage correct, for models in which both base rate and selection ratio, plus sample size and reliability of predictors, were manipulated. Based on Monte Carlo simulations, they found that ϕ_{p} was least affected by the base rate, followed by τ_{p} and then λ_{p}, for models that included continuous predictors. For models that involved only dichotomous predictors, however, τ_{p} was less sensitive than ϕ_{p} to changes in the base rate (and RIOC often could not be calculated at all, for reasons noted previously). Soderstrom and Leitner concluded that across a broad range of conditions, ϕ_{p} and τ_{p} were the most appropriate choices as indices of predictive efficiency. Taken together, these results reinforce the suggestion that selection of the index should be consistent with the nature of the model and, in particular, that τ_{p} is most appropriate for classification models and ϕ_{p} is most appropriate for selection models. These indices also may be applied to prediction tables generated by procedures other than logistic regression, and are not limited to dichotomous variables; they are applicable to any prediction table in which correct predictions can be distinguished from incorrect predictions. Unfortunately, none of the aforementioned indices of predictive efficiency is routinely available at present in such widely used logistic regression software as SPSS LOGISTIC REGRESSION or SAS PROC LOGISTIC.

2.4. Examples: Assessing the Adequacy of Logistic Regression Models

One reason for the lack of consensus about indices of predictive efficiency may be the fact that researchers are more often interested in the goodness of fit of the model (in a broad sense, as indicated by G_M and R_L^2) than in the accuracy of prediction or classification of the model, as indicated by the classification table and indices such as λ_p, τ_p, and ϕ_p. Especially for theory testing, goodness of fit is simply more important than accuracy of classification. The amount of space devoted to accuracy of prediction in this monograph reflects the relative lack of development in this area, rather than its importance, compared to the assessment of goodness of fit. Often the two approaches, goodness of fit and accuracy of prediction, will produce consistent results. It is entirely possible, however, to have a model that fits well, but does a poor job predicting category membership.

Figures 2.2 and 2.3 illustrate how indices of goodness of fit and predictive efficiency may lead to very different substantive conclusions. In Figure 2.2, hypothetical data are presented for a single dependent variable, TRUE, and a single predictor, P1. The standard output has been edited to include R_L^2, R^2, λ_p, and τ_p. For the 40 cases analyzed in Figure 2.2, the model fits well. $G_M = \text{model}\chi^2 = 20.123$ and is statistically significant (significance $= p = .0000$), leading us to reject the null hypothesis that the independent variable, P1, is not related to the dependent variable, TRUE. $R_L^2 = .363$, suggesting a moderate association between TRUE and P1. The binomial d is the same for both λ_p and τ_p (50% expected error for both): $d = 5.060$, with statistical significance $p = .000$. Both τ_p and λ_p are equal to .80, indicating that the independent variable allows us to classify the cases (into the categories of the dependent variable) with a very high degree of accuracy, as reflected in the classification table. Overall, the accuracy of prediction is considerably higher than the ability of the model to predict the probability, $P(Y_j = 1)$. The plot of observed groups and predicted probabilities at the bottom of Figure 2.2 indicates that the predicted probabilities are sometimes very high and sometimes close to .5, the cutoff for classification into $Y = 1$ or $Y = 0$. Accuracy of prediction is very high, even for cases whose predicted probability of belonging in $(Y = 1)$ is close to .5.

In Figure 2.3, G_M is again statistically significant, and both R_L^2 and R^2 indicate a moderately strong relationship between the dependent variable, TRUE, and the new predictor, P2. However, τ_p and λ_p indicate no more than a weak relationship between the observed and predicted classification of the cases, and the binomial $d = .632$, with statistical significance $p = .264$ (one-tailed), suggests that the classification on the dependent variable is not related to the values of



Classification Table(a)

			Predicted		
			TRUE		Percentage Correct
	Observed		0	1	
Step 0	TRUE	0	18	2	90.0
		1	2	18	90.0
	Overall Percentage				90.0

a The cut value is .500

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	20.123	1	.000
	Block	20.123	1	.000
	Model	20.123	1	.000

R_L^2 = .363
R^2 = .408
Tau-p = .80
Lambda-p = .80

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	35.329	.395	.527

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	P1	8.203	2.894	8.035	1	.005	3653.109
	Constant	-4.102	1.505	7.432	1	.006	.017

a Variable(s) entered on step 1: P1.

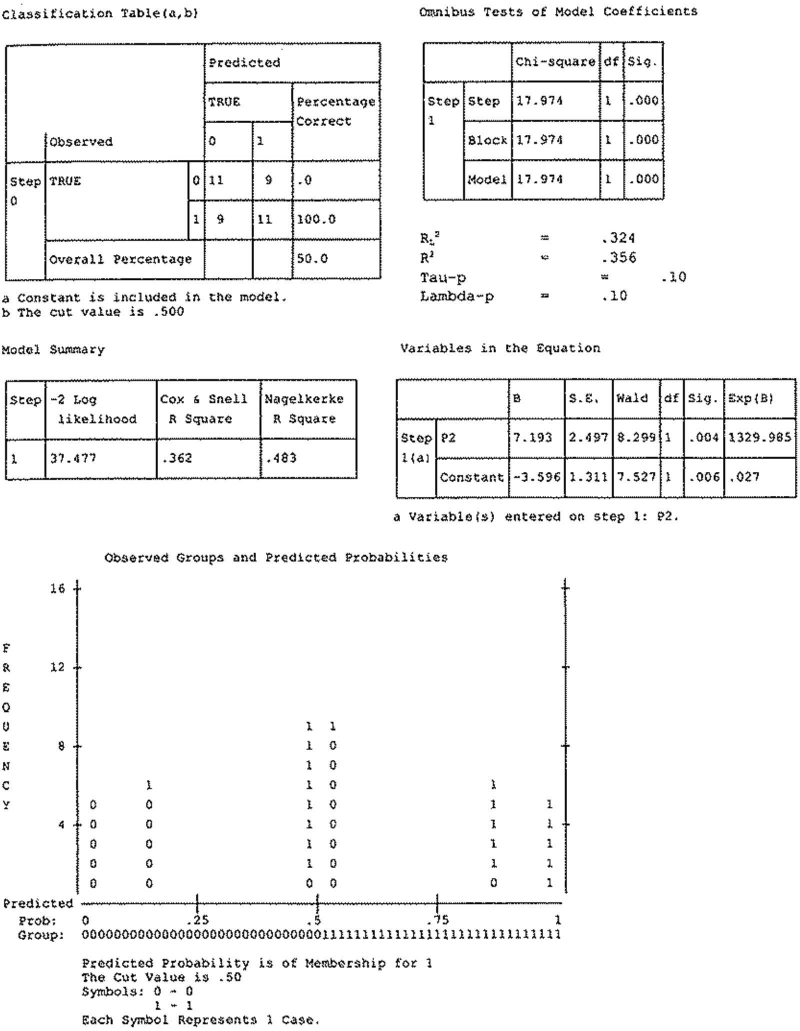
```
              Observed Groups and Predicted Probabilities

      16 +                                                            +
         |                                                            |
         |                                                            |
F        |                                                            |
R     12 +                                                            +
E        |                                                            |
Q        |                                                            |
U        |                            1  1                            |
E      8 +                            0  1                            +
N        |                            0  1                            |
C        |       1                    0  1                     1      |
Y        |  0    0                    0  1                     1    1 |
       4 +  0    0                    0  1                     1    1 +
         |  0    0                    0  1                     1    1 |
         |  0    0                    0  1                     1    1 |
         |  0    0                    0  0                     0    1 |
Predicted ---------------+--------------+--------------+--------------
  Prob:   0             .25            .5             .75            1
  Group:  000000000000000000000000000000111111111111111111111111111111

          Predicted Probability is of Membership for 1
          The Cut Value is .50
          Symbols: 0 - 0
                   1 - 1
          Each Symbol Represents 1 Case.
```

Figure 2.2. Logistic Regression Output for Hypothetical "Good Prediction" Data (Predicted Probability is of Membership)

the independent variable. The plot of observed groups and predicted probabilities indicates why. Now, instead of accurate prediction when the predicted probability is close to .5, predictions close to .5 are nearly all inaccurate. For 26 of the 40 cases, the predicted probabilities are the same for P1 and P2. For the other 14 cases, the predicted



Classification Table(a,b)

			Predicted		
			TRUE		Percentage Correct
	Observed		0	1	
Step 0	TRUE	0	11	9	.0
		1	9	11	100.0
	Overall Percentage				50.0

a Constant is included in the model.
b The cut value is .500

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	17.974	1	.000
	Block	17.974	1	.000
	Model	17.974	1	.000

R_L^2 = .324
R^2 = .356
Tau-p = .10
Lambda-p = .10

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	37.477	.362	.483

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	P2	7.193	2.497	8.299	1	.004	1329.985
	Constant	-3.596	1.311	7.527	1	.006	.027

a Variable(s) entered on step 1: P2.

```
              Observed Groups and Predicted Probabilities

       16 +                                                            +
          |                                                            |
          |                                                            |
F         |                                                            |
R      12 +                                                            +
E         |                                                            |
Q         |                                                            |
U         |                            1  1                            |
E       8 +                            1  0                            +
N         |                            1  0                            |
C         |        1                   1  0                   1        |
Y         | 0      0                   1  0                   1      1 |
        4 + 0      0                   1  0                   1      1 +
          | 0      0                   1  0                   1      1 |
          | 0      0                   1  0                   1      1 |
          | 0      0                   0  0                   0      1 |
Predicted  ---------------+--------------+--------------+--------------
  Prob:    0             .25            .5             .75            1
  Group:   000000000000000000000000000000111111111111111111111111111111

           Predicted Probability is of Membership for 1
           The Cut Value is .50
           Symbols: 0 - 0
                    1 - 1
           Each Symbol Represents 1 Case.
```

Figure 2.3. Logistic Regression Output for Hypothetical "Poor Prediction" Data

probability changed by .02, either from .49 to .51 or from .51 to .49. This had little impact on the overall goodness of fit of the model as measured by R_L^2, R^2, or η^2 and G_M, but it had a tremendous impact on the indices of predictive efficiency.

Figure 1.4, first discussed in Chapter 1, presented the results of a bivariate logistic regression analysis for real data, the relationship between exposure to delinquent friends (EDF5) and prevalence of marijuana use (PMRJ5). From Figure 1.4, we can reject the null hypothesis that EDF5 is unrelated to PMRJ5, based on G_{M} (significance = .0000). Notice that the Hosmer and Lemeshow goodness-of-fit measure is not statistically significant, indicating that the model with only EDF5 as a predictor fits the data well. The prediction table appears to indicate fairly good accuracy of prediction, but we need to calculate λ_{p} and/or τ_{p} to get a quantitative estimate of how well the cases are classified by the model. Finally, notice that the predicted values from the model have been saved as a new variable, LPEPMRJ5 (logistic regression prediction from EDF5 of PMRJ5). This permits us to use a separate analysis of variance or bivariate regression routine to calculate R^2.

From the data in Figure 1.4, where $G_{\mathrm{M}} = 85.359$ (the Model Chi-Square in the Omnibus Tests block) and $D_{\mathrm{M}} = 213.947$ (the −2 likelihood in the Model Summary block), $R_L^2 = G_{\mathrm{M}}/(G_m + D_{\mathrm{M}}) = (85.359)/(85.359 + 213.947) = .285$. Lambda-p is equal to the number of cases in the smaller observed category ($Y = 1$: $37 + 44 = 81$) minus the number of cases incorrectly predicted by the model ($37 + 14 = 51$), divided by the number of cases in the smaller category, so $\lambda_{\mathrm{p}} = (81 - 51)/81 = .370$. This is a moderately strong reduction in the error of *prediction*. Tau-p is a bit more complicated to calculate. First find the sum of the cases in each category for the observed value of Y: for $Y = 0$, $n_{Y=0} = 150$; for $Y = 1$, $n_{Y=1} = 81$. For a dichotomous dependent variable, the expected number of errors is the product of the two sums, divided by the total number of cases (231) and multiplied by 2 (because we expect the same number of errors in each category for a dichotomous variable), that is, $(2)(150)(81)/231 = 105.2$. Tau-p is the expected number of errors, minus the actual number of errors (51), divided by the expected number of errors: $\tau_{\mathrm{p}} = (105.2 - 51)/105.2 = .515$. This indicates that the model reduces the error of *classification* of cases as users or nonusers of marijuana by over half.

We can use the expected errors for λ_{p} and τ_{p} to calculate the binomial d statistic for each measure. For λ_{p}, the expected number of errors is 81, which corresponds to a proportion of $81/231 = .351$, and the observed number of errors is 51, which corresponds to a proportion of $51/231 = .221$. Therefore, $d = (P_{\mathrm{e}} - p_{\mathrm{e}})/\sqrt{P_{\mathrm{e}}(1 - P_{\mathrm{e}})/N} =$

$(.351 - .221)/\sqrt{(.351)(.649)/231} = 4.140$, with statistical significance $p = .000$. For τ_p, without going into as much detail, $d = 6.996$, with statistical significance $p = .000$. Finally, to calculate R^2 or η^2, we must use the results of a bivariate regression or an analysis of variance routine.

We can compare the results of the logistic regression analysis in Figure 1.4 directly with the results of the linear regression analysis, with the same variables, in Part C of Figure 1.1. In particular, the explained variance for PMRJ5 in the logistic regression model ($R^2 = .34$) is actually slightly higher than the explained variance for PMRJ5 in the linear regression model ($R^2 = .32$; from Part C of Figure 1.1).[13] This occurs despite the fact that the linear regression model tries to maximize R^2 (by minimizing the sum of the squared errors), whereas the logistic regression model does not. It appears that the logistic regression model fits the data well, indicates a moderately strong relationship between the predictor and the dependent variable, and does a fairly good job of predicting the classification of the cases.

2.5. Conclusion: Summary Measures for Evaluating the Logistic Regression Model

In linear regression, we use the F statistic and R^2 to test statistical significance and substantive significance, respectively, of the relationship between the dependent variable and the independent variables. Both are based on the total and error sums of squares, SST and SSE. In logistic regression, if our principal concern is how well the model fits the data (for example, in the context of theory testing), we use G_M and R_L^2, based on -2LL, to test for statistical and substantive significance. If our concern is less with the overall fit of the model and more with the accuracy with which the model predicts actual category membership on the dependent variable, the binomial d and one of the three indices of predictive efficiency (λ_p, τ_p, or ϕ_p) are used to assess the statistical and substantive significance of the model. Because λ_p, τ_p, and ϕ_p are not provided in logistic regression routines, they must be computed by hand.

3. INTERPRETING THE LOGISTIC REGRESSION COEFFICIENTS

In linear regression analysis, we evaluate the contribution of each independent variable to the model by testing for its statistical significance and then examining the substantive significance of its effect on the dependent variable. Statistical significance is evaluated using an F or t statistic to produce a probability (p) that we would find this strong a relationship in a sample this large if there really were no relationship between the independent variable and the dependent variable. Substantive significance may be evaluated in one of several ways. We may examine the unstandardized regression coefficient to see whether the change in the dependent variable associated with a given amount of change in the independent variable is large enough to be concerned about. (The words "are associated with" are used here in preference to language that would imply a causal relationship, because the relationships described here are definitely predictive and may, but need not, be causal in nature.) To apply this test, we must have some idea beforehand how big a change needs to be before we are concerned with it, and equivalently how big a change we are willing to ignore. Unstandardized regression coefficients are especially useful for evaluating the practical impact of one variable on another and for comparing the effects of the same variable in different samples.

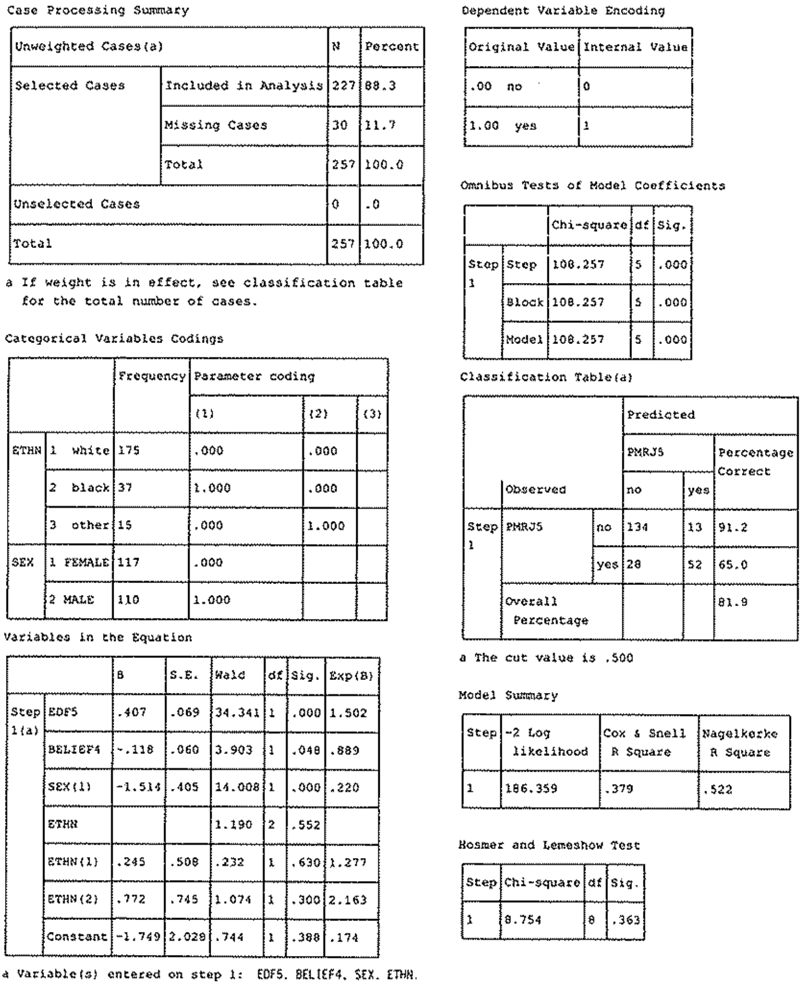
Alternatively, especially when there are no clear criteria for deciding "how big is big" and when some variables are not measured in natural units of measurement (feet, pounds, dollars), but are instead scale scores (an example of this would be the variable EDF5, exposure to delinquent friends), we may focus on a *standardized* regression coefficient, which indicates how many *standard deviations* a dependent variable changes in response to a 1 standard deviation change in the independent variable. Use of standardized coefficients is especially appropriate for theory testing and when the focus is on comparing the effects of different variables for the same sample.

For some purposes, stepwise methods are used to evaluate the contribution of variables to the regression equation. This is especially the case when we test for nonlinearity (for example, by including quadratic terms) or nonadditivity (by including interaction terms) in the regression equation. The decision about whether the inclusion of nonlinear or nonadditive terms is justified is typically based on the magnitude and statistical significance of the change in the explained variance, R^2. Stepwise methods are also used in exploratory analysis, when we are more concerned with theory development than theory testing. Such research may occur in the early stages of the study of a phenomenon, when neither theory nor knowledge about correlates of the phenomenon is well developed. Criteria for stepwise inclusion or removal of variables for a model generally involve tests that are similar to but less restrictive than the tests used in theory testing.

3.1. Statistical Significance in Logistic Regression Analysis

Several methods have been used to evaluate the statistical significance of the contribution of an independent variable to the explanation of a dependent variable. One stands out as being clearly the best, in the sense of being the most accurate: the likelihood ratio test. In the likelihood ratio test, the logistic regression model is calculated with and without the variable being tested. The likelihood ratio test statistic is equal to G_{M} for the model with the variable minus G_{M} for the model without the variable. The result, which we can call G_1 when we test X_1, G_2 when we test X_2, and G_k when we test X_k, has a χ^2 distribution with degrees of freedom equal to the degrees of freedom in the model with X minus the degrees of freedom in the model without X. For example, if we designate G_{M1} to represent the model χ^2 with X_k in the model and designate G_{M2} to represent the model χ^2 with X_k not in the model, $G_k = G_{\mathrm{M1}} - G_{\mathrm{M2}}$, and if X is a continuous, interval, or ratio variable, then G_k has 1 degree of freedom.

The only drawback to the use of the likelihood ratio statistic is that it requires more time to compute than alternative tests for statistical significance. If you are paying for every second on a mainframe computer, this may be a serious concern, but for many users with access to relatively fast personal computers and workstations, this is irrelevant except for very large samples. Nonetheless, statistical packages are often written to use a less computationally intensive alternative to the likelihood test, the Wald statistic, to test for the statistical significance of individual coefficients. In Figure 1.4, the Wald statistic appears following the coefficient (B) and its standard error (S.E.). The Wald statistic may be calculated as $W_k^2 = [b_k/(\text{S.E. of } b_k)]^2$, in which case it is asymptotically distributed as a χ^2 distribution or as $W_k = b_k/(\text{S.E. of } b_k)$, in which case it follows a standard normal distribution (Hosmer & Lemeshow, 1989, p. 31; SAS, 1989, p. 1097; SPSS, 1991, pp. 140–141) and its formula parallels the formula for the t ratio for coefficients in linear regression. The disadvantage of the Wald statistic is that for large b, the estimated standard error is inflated, resulting in failure to reject the null hypothesis when the null hypothesis is false.[14]



Case Processing Summary

Unweighted Cases(a)		N	Percent
Selected Cases	Included in Analysis	227	88.3
	Missing Cases	30	11.7
	Total	257	100.0
Unselected Cases		0	.0
Total		257	100.0

a If weight is in effect, see classification table for the total number of cases.

Categorical Variables Codings

		Frequency	Parameter coding		
			(1)	(2)	(3)
ETHN	1 white	175	.000	.000	
	2 black	37	1.000	.000	
	3 other	15	.000	1.000	
SEX	1 FEMALE	117	.000		
	2 MALE	110	1.000		

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	EDF5	.407	.069	34.341	1	.000	1.502
	BELIEF4	-.118	.060	3.903	1	.048	.889
	SEX(1)	-1.514	.405	14.008	1	.000	.220
	ETHN			1.190	2	.552	
	ETHN(1)	.245	.508	.232	1	.630	1.277
	ETHN(2)	.772	.745	1.074	1	.300	2.163
	Constant	-1.749	2.029	.744	1	.388	.174

a Variable(s) entered on step 1: EDF5, BELIEF4, SEX, ETHN.

Dependent Variable Encoding

Original Value	Internal Value
.00 no	0
1.00 yes	1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	108.257	5	.000
	Block	108.257	5	.000
	Model	108.257	5	.000

Classification Table(a)

			Predicted		
			PMRJ5		Percentage Correct
	Observed		no	yes	
Step 1	PMRJ5	no	134	13	91.2
		yes	28	52	65.0
	Overall Percentage				81.9

a The cut value is .500

Model Summary

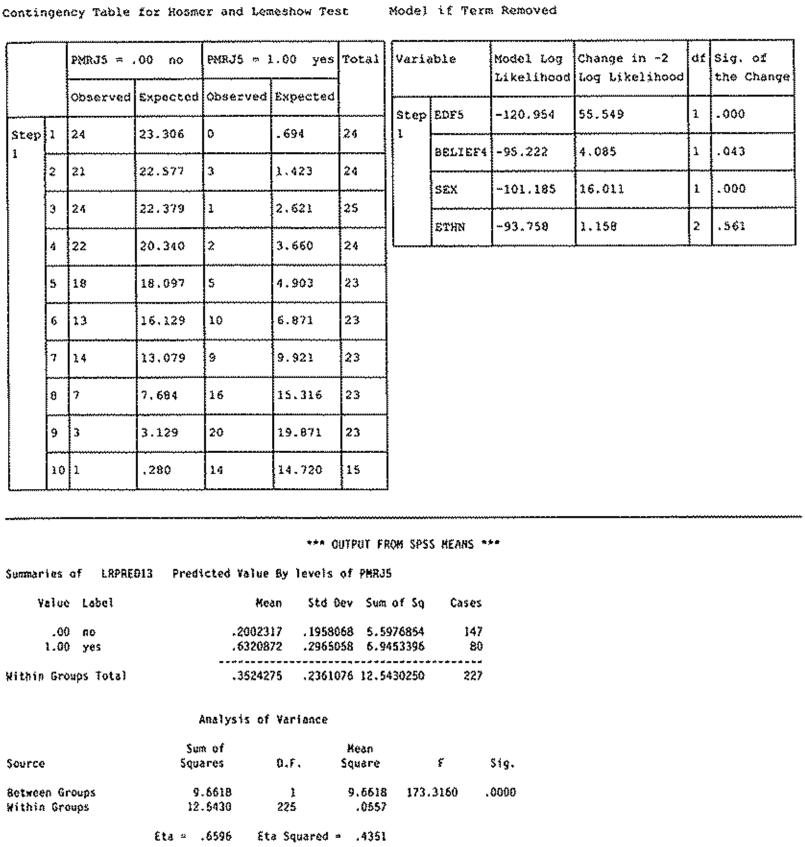
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	186.359	.379	.522

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8.754	8	.363

Figure 3.1. SPSS LOGISTIC REGRESSION Output

Figure 3.1 presents SPSS output for a dependent variable with four predictors. PMRJ5, the prevalence of marijuana use, is again the dependent variable and EDF5, exposure to delinquent friends, is again included as a predictor. BELIEF4 is a scale that measures how wrong (very wrong, wrong, a little wrong, not wrong at all) the respondent believes it is to commit each of several illegal acts (assault, theft,



Contingency Table for Hosmer and Lemeshow Test

		PMRJ5 = .00 no		PMRJ5 = 1.00 yes		Total
		Observed	Expected	Observed	Expected	
Step 1	1	24	23.306	0	.694	24
	2	21	22.577	3	1.423	24
	3	24	22.379	1	2.621	25
	4	22	20.340	2	3.660	24
	5	18	18.097	5	4.903	23
	6	13	16.129	10	6.871	23
	7	14	13.079	9	9.921	23
	8	7	7.684	16	15.316	23
	9	3	3.129	20	19.871	23
	10	1	.280	14	14.720	15

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	EDF5	-120.954	55.549	1	.000
	BELIEF4	-95.222	4.085	1	.043
	SEX	-101.185	16.011	1	.000
	ETHN	-93.758	1.158	2	.561

```
                              *** OUTPUT FROM SPSS MEANS ***

Summaries of   LRPRED13   Predicted Value By levels of PMRJ5

    Value  Label                  Mean     Std Dev  Sum of Sq     Cases

      .00  no                 .2002317   .1958068  5.5976854       147
     1.00  yes                .6320872   .2965058  6.9453396        80
                          ------------------------------------------
Within Groups Total           .3524275   .2361076 12.5430250       227


                        Analysis of Variance

                         Sum of                  Mean
Source                   Squares       D.F.      Square        F         Sig.

Between Groups            9.6618          1      9.6618   173.3160      .0000
Within Groups            12.6430        225       .0557

                    Eta =  .6596    Eta Squared =  .4351
```

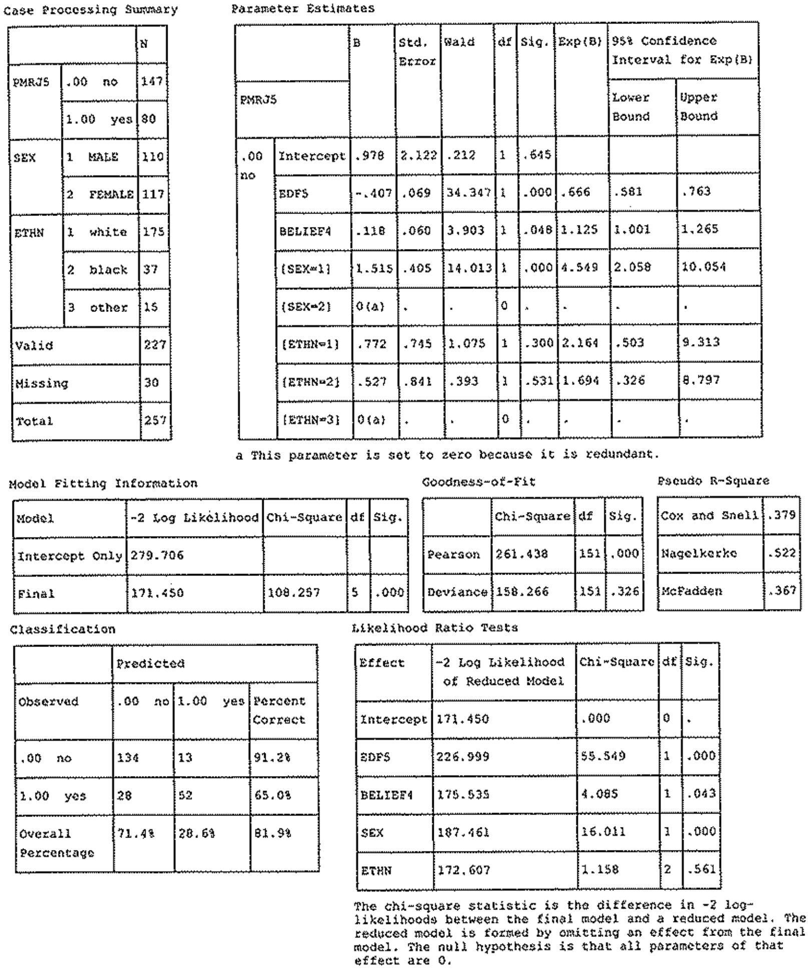
Figure 3.1. (Continued)

selling hard drugs, etc.), parallel to the items used to construct EDF5. BELIEF4 is measured immediately prior to the period for which data were collected on prevalence of marijuana use. SEX is coded 0 for females and 1 for males.

The coding for ETHN is given in the Categorical Variables Codings Table in Figure 3.1. The first coefficient for ETHN corresponds to being African American ("black") and the second corresponds to being other than non-Hispanic European American or African American ("other"). The numbers in the columns, (1), (2) or (3), correspond to the number of the coefficient that is set to 1 for individuals falling into each row. Thus, the first coefficient for ETHN is multiplied by 1 for African Americans, and by 0 otherwise, and the second coefficient is multiplied by 1 for other ethnic groups and by 0 otherwise. This is an example of the use of a set of "dummy" or design variables in logistic regression to represent a single categorical variable, and is parallel to the use of dummy variables or design variables in linear regression (Hardy, 1993; Lewis-Beck, 1980).

In the section of Figure 3.1 labeled Variables in the Equation, we find logistic regression coefficients, standard errors, Wald statistics (W_k^2), the degrees of freedom (df) associated with each variable, and the statistical significance of the Wald statistic. From Figure 3.1, it appears that EDF5, BELIEF4, and SEX have statistically significant effects on PMRJ5. For ETHN, the Wald statistic is computed for the variable as a whole, and also separately for each of the coefficients that correspond to the separate categories of ethnicity. The effect of ETHN on PMRJ5 does not appear to be statistically significant, and neither does the intercept (constant). The Hosmer–Lemeshow test (here the contingency table for the Hosmer–Lemeshow test is also included) indicates a good fit for the model. Toward the end of Figure 3.1, just before the output from SPSS MEANS, under the heading Model if Term Removed, likelihood ratio statistics were obtained for the variables in the model.[15] Substantively, the conclusions remain the same and the significance levels are very similar for the Wald and likelihood ratio statistics. As noted earlier, the Cox–Snell and Nagelkerke R^2 measures are questionable measures, so they are not discussed here.

Figure 3.2 provides parallel output from SPSS NOMREG. The NOMREG output is more concise and reveals some of the same information in the output from LOGISTIC REGRESSION. The model χ^2 (G_M) is the same, as is the (final) -2LL (D_M) for the model (compare the Model Summary table in Figure 3.1 and the Model Fitting Information table in Figure 3.2), and the Intercept Only or initial -2LL (D_0) is also provided. The Cox–Snell and Nagelkerke measures are the same in the two tables, but NOMREG also provides the McFadden pseudo R-square (R_L^2). The Likelihood Ratio Tests table in Figure 3.2 provides the same information as the Model if Term Removed table in Figure 3.1. In place of the Hosmer–Lemeshow test for model fit, NOMREG provides the Pearson and deviance χ^2 statistics, adjusted to reflect the number of covariate patterns rather than the number of cases. Of the two, the deviance χ^2 is generally regarded



Case Processing Summary

		N
PMRJ5	.00 no	147
	1.00 yes	80
SEX	1 MALE	110
	2 FEMALE	117
ETHN	1 white	175
	2 black	37
	3 other	15
Valid		227
Missing		30
Total		257

Parameter Estimates

		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
PMRJ5								Lower Bound	Upper Bound
.00 no	Intercept	.978	2.122	.212	1	.645			
	EDF5	-.407	.069	34.347	1	.000	.666	.581	.763
	BELIEF4	.118	.060	3.903	1	.048	1.125	1.001	1.265
	[SEX=1]	1.515	.405	14.013	1	.000	4.549	2.058	10.054
	[SEX=2]	0(a)	.	.	0	.	.	.	.
	[ETHN=1]	.772	.745	1.075	1	.300	2.164	.503	9.313
	[ETHN=2]	.527	.841	.393	1	.531	1.694	.326	8.797
	[ETHN=3]	0(a)	.	.	0	.	.	.	.

a This parameter is set to zero because it is redundant.

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	279.706			
Final	171.450	108.257	5	.000

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	261.438	151	.000
Deviance	158.266	151	.326

Pseudo R-Square

Cox and Snell	.379
Nagelkerke	.522
McFadden	.367

Classification

	Predicted		
Observed	.00 no	1.00 yes	Percent Correct
.00 no	134	13	91.2%
1.00 yes	28	52	65.0%
Overall Percentage	71.4%	28.6%	81.9%

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	171.450	.000	0	.
EDF5	226.999	55.549	1	.000
BELIEF4	175.535	4.085	1	.043
SEX	187.461	16.011	1	.000
ETHN	172.607	1.158	2	.561

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Figure 3.2. Logistic Regression Output From SPSS NOMREG

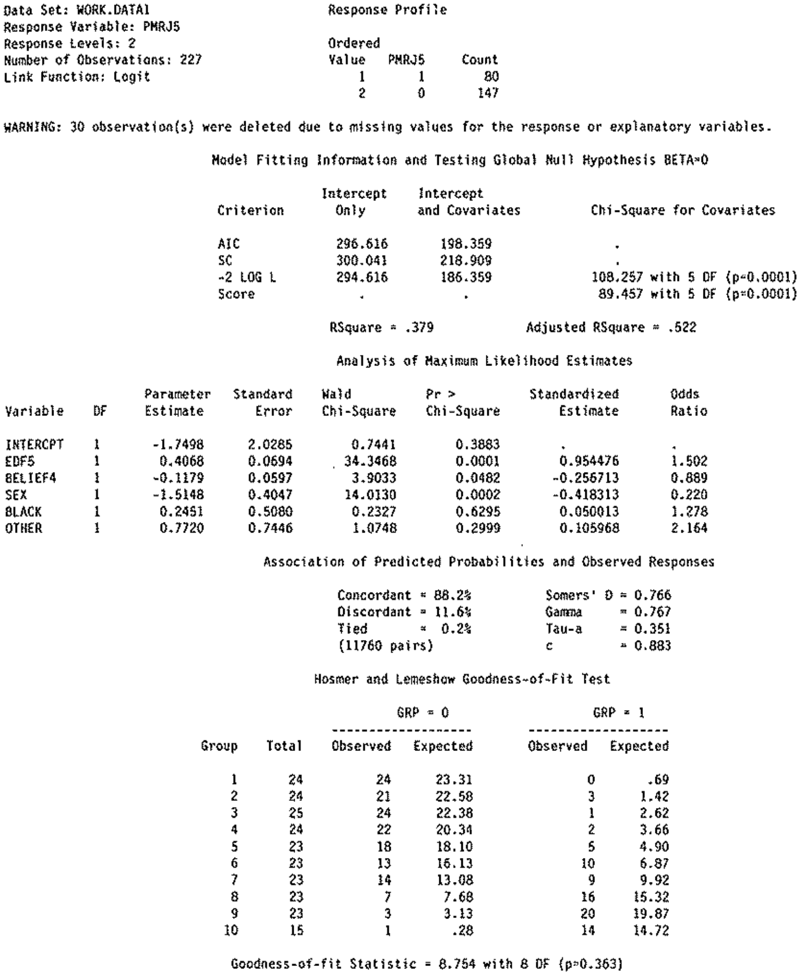
to be more informative than the Pearson χ^2 in logistic regression, and the deviance χ^2 indicates a good model fit, similar to the Hosmer–Lemeshow test in Figure 3.1. The classification table is also identical.

The one striking difference between Figures 3.2 and 3.1 is in the coefficients in the Parameter Estimates table in Figure 3.2. LOGISTIC REGRESSION predicts to the second category of the dependent variable (thus setting the first category as the reference category) and allows us to specify which category is the reference category for each categorical predictor, and in Figure 3.1, the first category was selected as the reference category. By default, NOMREG uses the last category as the reference category for both the dependent and independent variables. As a result, some of the coefficients seem different, but they really tell the same story as the coefficients in Figure 3.1. Individuals with higher exposure to delinquent friends are *less* likely to be *nonusers*; individuals with strong beliefs that it is wrong to violate the law are *more* likely to be *nonusers*; and males are more likely to be *nonusers* (note that it is the first category of the dependent variable that is being predicted here) according to the results in Figure 3.2. In Figure 3.2, it is "white" instead of "other" that is the reference category, but still neither of the two logistic regression coefficients is statistically significant.

Figure 3.3 provides partial output from SAS for the same model that was estimated in Figure 3.1. One difference between the SAS output in Figure 3.3 and the SPSS output in Figure 3.1 is the treatment of the categorical variable ETHN, which in SAS must be separated into design variables before it is entered into the analysis, because SAS assumes that the independent variables in PROC LOGISTIC have true numeric values (SAS, 1989, p. 1079). Also, there is no test for ETHN as a single variable; each of the design variables is tested separately. Such a test could be performed by analyzing separate models, one with and one without ETHN as a predictor. Otherwise, however, the parameter estimates, standard errors, Wald statistics, p values, Hosmer–Lemeshow test, R^2_{M} (RSquare in Figure 3.3; Cox–Snell in SPSS), R^2_{N} (Adjusted RSquare in Figure 3.3; Nagelkerke in SPSS), and −2 Log likelihood statistics are practically identical to those produced by SPSS LOGISTIC REGRESSION.

3.2. Interpreting Unstandardized Logistic Regression Coefficients

Figure 1.4 provided the results of a bivariate logistic regression analysis. From Figure 1.4, we obtained the equation logit(PMRJ5) = .407(EDF5) − 5.487. When EDF5 is at its maximum observed value (29), this becomes logit(PMRJ5) = .407(29) − 5.487 = 6.316. If EDF5 has its minimum observed value (8) it becomes logit(PMRJ5) = .407(8) − 5.487 = −2.231. Translating the logits into probabilities, the probability of marijuana use for individuals whose score on the



```
Data Set: WORK.DATA1                        Response Profile
Response Variable: PMRJ5
Response Levels: 2                          Ordered
Number of Observations: 227                 Value   PMRJ5     Count
Link Function: Logit                            1       1        80
                                                2       0       147

WARNING: 30 observation(s) were deleted due to missing values for the response or explanatory variables.

                            Model Fitting Information and Testing Global Null Hypothesis BETA=0

                             Criterion     Intercept    Intercept
                                            Only        and Covariates          Chi-Square for Covariates

                             AIC            296.616        198.359                  .
                             SC             300.041        218.909                  .
                             -2 LOG L       294.616        186.359              108.257 with 5 DF (p=0.0001)
                             Score             .              .                  89.457 with 5 DF (p=0.0001)

                                            RSquare = .379            Adjusted RSquare = .522

                                             Analysis of Maximum Likelihood Estimates

                   Parameter   Standard    Wald           Pr >           Standardized        Odds
Variable    DF     Estimate       Error    Chi-Square     Chi-Square         Estimate        Ratio

INTERCPT    1       -1.7498      2.0285        0.7441         0.3883          .               .
EDF5        1        0.4068      0.0694       34.3468         0.0001         0.954476        1.502
BELIEF4     1       -0.1179      0.0597        3.9033         0.0482        -0.256713        0.889
SEX         1       -1.5148      0.4047       14.0130         0.0002        -0.418313        0.220
BLACK       1        0.2451      0.5080        0.2327         0.6295         0.050013        1.278
OTHER       1        0.7720      0.7446        1.0748         0.2999         0.105968        2.164

                                   Association of Predicted Probabilities and Observed Responses

                                             Concordant = 88.2%          Somers' D = 0.766
                                             Discordant = 11.6%          Gamma     = 0.767
                                             Tied       =  0.2%          Tau-a     = 0.351
                                             (11760 pairs)               c         = 0.883

                                          Hosmer and Lemeshow Goodness-of-Fit Test

                                                      GRP = 0                    GRP = 1
                                             -------------------        -------------------
                           Group    Total    Observed   Expected        Observed   Expected

                               1       24          24      23.31               0        .69
                               2       24          21      22.58               3       1.42
                               3       25          24      22.38               1       2.62
                               4       24          22      20.34               2       3.66
                               5       23          18      18.10               5       4.90
                               6       23          13      16.13              10       6.87
                               7       23          14      13.08               9       9.92
                               8       23           7       7.68              16      15.32
                               9       23           3       3.13              20      19.87
                              10       15           1        .28              14      14.72

                               Goodness-of-fit Statistic = 8.754 with 8 DF (p=0.363)
```

Figure 3.3. SAS PROC LOGISTIC Output

exposure scale is 29 becomes $e^{6.316}/(1 + e^{6.316})$=.998; for individuals whose score on the exposure scale is 8, it becomes $e^{-2.231}/(1 + e^{-2.231})$=.097. For individuals with the highest levels of exposure to delinquent friends, marijuana use is almost, but not quite, certain. For individuals with the lowest levels of exposure, the relative frequency of marijuana use is less than 10%, low, but far from indicating that marijuana use never occurs among these individuals. At the mean level of exposure, logit(PMRJ5) = .407(12) − 5.487 = −.603, and the probability of using marijuana is $e^{-.603}/(1 + e^{-.603}) = .354$, which is approximately equal to the unconditional probability of marijuana use ($P = .357$) for this 16-year-old sample.

Like the linear regression coefficient, the logistic regression coefficient can be interpreted as the change in the dependent variable, logit(Y), associated with a one-unit change in the independent variable. The change in P(Y = 1), however, is *not* a linear function of the independent variables. The slope of the curve varies, depending on the value of the independent variables. It is possible to calculate the slope of the curve for different pairs of points by examining the change in P(Y = 1) between those points. For example, going from EDF5 = 8 to EDF5 = 9 results in a change in probability from .097 to .101, indicating a slope of .004. A change from EDF5 = 28 to EDF5 = 29 is associated with a change from P(Y = 1) = .997 to .998, or a slope of about .001. Between EDF5 = 12 and EDF5 = 13, the probability of marijuana use changes from .354 to .451, a slope of .097, which is many times larger than the changes that result from one-unit changes at very high or very low values of EDF5.

The interpretation of the logistic regression coefficient is similar in models with several independent variables. The equation for the relationship between prevalence of marijuana use and the predictors in Figure 3.1 is logit(PMRJ5) = .407(EDF5) − .118(BELIEF4) − 1.514(SEX) + .245(BLACK) + .772(OTHER) − 1.749, where BLACK and OTHER are the descriptive labels associated with ETHN(1) and ETHN(2) in Figure 3.1. Turning to the individual coefficients, each one-unit increase in EDF5 is associated with an increase of .407 in logit(PMRJ5). Each one-unit increase in BELIEF4 is associated with a decrease of .118 in logit(PMRJ5). Being male reduces the logit of PMRJ5 by 1.514 (remember, in this sample, males have lower marijuana use than females). The effects of ethnicity are not statistically significant.[16]

Predictions for individual cases may be obtained by replacing the variables in the equation with their values for specific cases. For example, for an African American female (BLACK = 1, OTHER = 0) with strong beliefs that it is wrong to violate the law (BELIEF4 = 25) and low levels of exposure to delinquent friends (EDF5 = 10), logit(PMRJ5) = .407(10) − .118(25) − 1.514(0) + .245(1) + .772(0) − 1.749 = −.384. This corresponds to a probability of marijuana use of

$e^{-.384}/(1+e^{-.384}) = .405$. Alternatively, for a non-Hispanic European American male (BLACK = 0, OTHER = 0) with moderate levels of both belief that it is wrong to violate the law (BELIEF4 = 20) and exposure to delinquent friends (EDF5 = 15), the equation becomes $\text{logit(PMRJ5)} = .407(15) - .118(20) - 1.514(1) + .245(0) + .772(0) - 1.749 = .482$. This corresponds to a probability of marijuana use of $e^{.482}/(1 + e^{.482}) = .618$.

3.3. Substantive Significance and Standardized Coefficients

What does a one-unit increase in exposure to delinquent friends really mean? Because exposure items are measured on a five-point scale, and belief items are measured on a four-point scale, and because the number of items is larger for exposure (8) than for belief (7), is a one-unit increase in belief really equivalent to a one-unit increase in exposure? Should we regard a one-unit change in belief (which has, in principle, a range of 7 to 28) as equivalent to a one-unit change in gender (which has a range of only one unit, 0 to 1)? These questions could be asked in the context of either linear regression, with frequency of marijuana use as a dependent variable, or logistic regression with prevalence of marijuana use as a dependent variable. When independent variables are measured in different units or on different scales and we want to compare the strength of the relationship between the dependent variable and different independent variables, we often use standardized regression coefficients in linear regression analysis. For the same reasons, we may want to consider using standardized coefficients in logistic regression analysis.

A standardized coefficient is a coefficient that has been calculated for variables measured in standard deviation units. A standardized coefficient indicates how many standard deviations of change in a dependent variable are associated with a 1 standard deviation increase in the independent variable. In linear regression, a standardized coefficient between a dependent variable Y and an independent variable X, b^*_{YX}, may be calculated from the unstandardized coefficient between Y and X, b_{YX}, and the standard deviations of the two variables, s_Y, and s_X: $b^*_{YX} = (b_{YX})(s_X)/(s_Y)$. Alternatively, standardizing both X and Y prior to regression by subtracting their respective means and dividing by their respective standard deviations to obtain $Z_Y = (Y - \overline{Y})/s_Y$ and $Z_X = (X - \overline{X})/s_X$ produces a standardized regression coefficient between Y and X.

For a variable that is approximately normally distributed, 99.9865% of all cases will lie in a range of 6 standard deviations (3 standard deviations on either side of the mean), and 99.999999713 will lie within a range of 10 standard deviations. Thus, a 1 standard deviation change in an independent variable typically means a change of about one-eighth of the range of its possible values (one-sixth in a small sample; one-tenth in a very large sample). According to Chebycheff's inequality theorem (Bohrnstedt & Knoke, 1994, pp. 82–83), for *any* distribution, even for a very nonnormal distribution, *at least* 93.75% of all cases will lie within 8 standard deviations of the mean and 96% will lie within 10 standard deviations. Thus a change of 1 standard deviation seems intuitively to be a large enough change that its effect should be felt (if the independent variable has any impact on the dependent variable), but not so large that a trivial relationship should appear to be substantial, even in a distribution that departs considerably from a normal distribution. By measuring the relationship of all of the independent variables to the dependent variables in common units (standard deviations, or about one-eighth of their range), the relative impact on the dependent variable of independent variables measured in different units can be directly compared.

In logistic regression analysis, the calculation of standardized coefficients is complicated by the fact that it is not the value of Y, but the probability that Y has one or the other of its possible values, that is predicted by the logistic regression equation. The actual dependent variable in logistic regression is not Y, but logit(Y), whose observed values of logit(0) $= -\infty$ and logit($+\infty$) $= +\infty$ do not permit the calculation of means or standard deviations. Although we cannot calculate the standard deviation directly for the observed values of logit(Y), we can calculate the standard deviation indirectly, using the predicted values of logit(Y) and the explained variance, R^2. Recall from Chapter 2 that R^2 = SSR/SST. Dividing both the numerator and the divisor by N($N - 1$ for a sample), we get R^2 = SSR/SST = (SSR/N)/(SST/N) $= s_{\hat{Y}}^2/s_Y^2$. Rearranging this equation to solve for s_Y^2 produces the equation $s_Y^2 = s_{\hat{Y}}^2/R^2$, and substituting logit(Y) for Y and logit($\widehat{Y}$) for ($\widehat{Y}$), we are able to calculate the variance of logit(Y) based on the standard deviation of the predicted values of logit(Y) and the explained variance. Because the standard deviation is the square root of the variance, we can estimate standardized logistic regression coefficients as

$$b_{YX}^* = (b_{YX})(s_X)/\sqrt{s_{\text{logit}(\widehat{Y})}/R^2} = (b_{YX})(s_X)(R)/s_{\text{logit}(\widehat{Y})}, \qquad [3.1]$$

where b^*_{YX} is the standardized logistic regression coefficient, b_{YX} is the unstandardized logistic regression coefficient, s_X is the standard deviation of the independent variable X, $s^2_{\text{logit}(\widehat{Y})}$ is the variance of $\text{logit}(\widehat{Y})$ [in other words, the variance of the estimated values of $\text{logit}(Y)$], $s_{\text{logit}(\widehat{Y})}$ is the standard deviation of $\text{logit}(\widehat{Y})$, and R^2 is the coefficient of determination.

To calculate standardized logistic regression coefficients with existing SAS and SPSS software, the following steps are necessary:

1. b: Calculate the logistic regression model to obtain the unstandardized logistic regression coefficient b. Save the predicted value of Y from the logistic regression model.
2. R: Use the predicted value of Y to calculate R^2, R, η^2, or η (because these measures convey the same information, it does not matter which one you calculate).
3. Use the predicted value of Y to calculate the predicted value of $\text{logit}(Y)$, using the equation $\text{logit}(\widehat{Y}) = \ln[\widehat{Y}/(1 - \widehat{Y})]$.
4. $s_{\text{logit}(\widehat{Y})}$: Calculate descriptive statistics for $\text{logit}(\widehat{Y})$, including the standard deviation.
5. s_X: If you have not already done so, calculate the standard deviations of all of the independent variables in the equation. Be sure that you calculate them only for the cases actually included in the model. (In other words, use listwise deletion of missing data when you calculate the descriptive statistics.)
6. Enter b, R (or η), s_X, and $s_{\text{logit}(\widehat{Y})}$ into Equation 3.1 to calculate b^*.

The interpretation of the standardized logistic regression coefficient, calculated as $b^* = bs_XR/s_Y$, is straightforward and closely parallels the interpretation of standardized coefficients in linear regression: a 1 standard deviation increase in X produces a b^* standard deviation change in $\text{logit}(Y)$. For the model in Figure 1.4, the standard deviation of EDF5, the standard deviation of $\text{logit}(\widehat{Y})$, and η were obtained separately. The standard deviation of EDF5 was 4.24, the standard deviation of $\text{logit}(\widehat{Y})$ was $s_{\text{logit}(\widehat{Y})} = 1.72$, $R = \eta = .5871$, and $b = .4068$. From Equation 3.1, $b^* = (.4068)(4.24)(.5871)/1.72 = .591$. In other words, a 1 standard deviation increase in EDF5 is associated with an increase of .591 standard deviations in logit(PMRJ5).

Table 3.1 summarizes the output from SAS PROC LOGISTIC and SPSS LOGISTIC REGRESSION, and adds measures of explained variation and predictive efficiency. The relationship between the

TABLE 3.1

Logistic Regression Analysis Results for Prevalence of Marijuana Use

Dependent Variable	*Association/ Predictive Efficiency*	*Independent Variable*	*Unstandardized Logistic Regression Coefficient (b)*	*Standard Error of b*	*Statistical Significance of b*	*Standardized Logistic Regression Coefficient*
PMRJ5	$G_M = 108.257$ $(p = .000)$	EDF5	.407	.069	.000	.531
		BELIEF4	−.118	.060	.048	−.143
	$R^2_L = .367$					
		SEX (male)	−1.514	.405	.000	−.233
	$R^2 = .435$					
		ETHN			.552	
	$\lambda_p = .488$	Black	.245	.508	.630	.028
		Other	.772	.745	.300	.059
	$\tau_p = .604$					
		Intercept	−1.749	2.028	.388	—

dependent variable and the independent variables is statistically significant: $G_{\mathrm{M}} = 108.257$ with 5 degrees of freedom, $p = .000$. Measures of the strength of association between the dependent variable and the independent variables, $R^2_{\mathrm{L}} = .367$ and $R^2 = \eta^2 = .435$ (the latter from Figure 3.1), indicate a moderately strong relationship between the dependent variable and its predictors. The indices of predictive efficiency also indicate a model that predicts well: $\lambda_{\mathrm{p}} = .488$ and $\tau_{\mathrm{p}} = .604$, both statistically significant at $p = .000$.

Comparison of Table 3.1 with Figure 3.3 reveals that the standardized coefficients in Table 3.1 do not match the "Standardized Estimate" provided by SAS in Figure 3.3. This is because SAS calculates the standardized estimate of the logistic regression coefficient as $b^*_{SAS} = (b)(s_X)/(\pi/\sqrt{3}) = (b)(s_X)/1.8138$. The quantity $\pi/\sqrt{3}$ is the standard deviation of the standard logistic distribution (just as 1 is the standard deviation of the standard normal distribution). The "standardized" coefficients provided by SAS are really partially, not fully, standardized; they do not take the actual distribution of Y or logit(Y) into account, but divide by the same constant regardless of the distribution of Y. Another alternative is to standardize only the independent variables. Both the SAS and independents-only approach to partial standardization produce the same ranking of effects of independent variables on the dependent variable as full standardization, but limited experience suggests that they are more likely than the fully standardized coefficient b^* to be greater than 1 or less than -1 even when there are no collinearity or other problems (see Chapter 4). The principal reasons to favor the use of the fully standardized coefficient are (a) construction and interpretation that directly parallel the standardized coefficients in linear regression and, correspondingly, (b) the ability to apply the same standards used to interpret standardized coefficients in linear regression to standardized coefficients in logistic regression.

If we naively evaluate the strength of the relationships of the independent variables to PMRJ5 based on the unstandardized logistic regression coefficients (or equivalently, based on odds ratios or probabilities), SEX appears to have the strongest effect, followed by EDF5 and BELIEF4. (ETHN is not statistically significant.) Based on the standardized coefficients, however, EDF5 appears to have the strongest effect (.531 in Table 3.1), followed by SEX ($-.233$) and then BELIEF4 ($-.143$). In other words, (a) a 1 standard deviation increase in EDF5 is associated with a .531 standard deviation increase in logit (PMRJ5); (b) a 1 standard deviation increase in BELIEF4 is associated with a .143 standard deviation decrease in logit (PMRJ5); and (c) a 1 standard deviation increase (becoming "more male") in SEX is associated with a .418 standard deviation decrease in logit (PMRJ5). Changes in ETHN, which is not statistically significant as a predictor of PMRJ5, are associated with changes of less than one-tenth of a standard deviation in logit (PMRJ5).

For SEX and ETHN, a 1 standard deviation increase is not as intuitively meaningful as the difference between males and females or between respondents from different ethnic backgrounds, as reflected in the unstandardized logistic regression coefficient. The real utility of the standardized logistic regression coefficient here is to compare the magnitude of the effects of the predictors by converting them to a common scale of measurement. In presenting substantive results, it may make sense to focus on standardized coefficients for unitless scales like EDF5 and BELIEF4, but unstandardized or exponentiated coefficients for categorical variables like ETHN and SEX (corresponding to realistic differences in ethnicity and gender), and perhaps for variables with natural units of measurement (inches, kilograms, dollars, number of occasions) as well.

3.4. Exponentiated Coefficients or Odds Ratios

In the last column of statistics in Figures 3.1 under Variables in the Equation, the third to last column in Figure 3.2 under Parameter Estimates, and the last column in Figure 3.3 under the heading Analysis of Maximum Likelihood Estimates, the *odds ratio* associated with each coefficient is presented, as Exp(B) in SPSS and as Odds Ratio in SAS. The odds ratio is the number by which we would multiply the odds of being a marijuana user (the probability divided by 1 minus the probability) for each one-unit increase in the independent variable. An odds ratio greater than 1 indicates that the odds of being a marijuana user increase when the independent variable increases; an odds ratio of less than 1 indicates that the odds of being a marijuana user decrease when the independent variable increases. For example, a one-unit increase in EDF5 results in a 50.2% increase in the odds of being a marijuana user (the odds of being a marijuana user is multiplied by 1.502). A one-unit increase in BELIEF4 decreases the odds of being a marijuana user by 11.2% (the odds of being a marijuana user is multiplied by .889, which is .112 less than 1).

It is important to emphasize that the odds ratio is not a separate measure of the relationship between the dependent variables and the independent variables. It contains the same information as the logistic regression coefficient or the probability. All that is different is the way in which the information is presented. In particular, the odds ratio *cannot* take the place of a standardized logistic regression coefficient for evaluating the strength of the influences of the independent variables on the dependent variable, relative to one another, because the odds ratio will provide exactly the same ordering, from strongest to weakest, as the unstandardized logistic regression coefficient, once all of the odds ratios are transformed to be greater than 1 (or all less than 1). The odds ratio provides no additional information; it just provides the same information as the logistic regression coefficient in a different way.

I have repeatedly seen the mistake of equating the *odds ratio* (a ratio of two odds) with a *risk ratio* (a ratio of two probabilities), sometimes with the justification that the two are "approximately" equal under certain fairly restrictive conditions (a base rate less than .10). In general, the use of an odds ratio to "represent" a risk ratio will overstate the strength of the relationship. An odds ratio of about .22 for males (see Figures 3.1 and 3.3) does not mean that the *risk* of marijuana use is only a little over one-fifth as high for males as for females or that the odds ratio of 4.5 for females (see Figure 3.2) indicates that the risk of marijuana use is nearly five times as high for females as for males. To compare the *relative risk* of marijuana use for males and females, it is necessary to use the model to calculate the probabilities for each, assuming values of the other predictors. For white males and females with average levels of EDF5 (12) and BELIEF4 (27), the respective probabilities are for females $e^{.407(12)-.118(27)-1.749}/(1 + e^{.407(12)-.118(27)-1.749}) = .487$ and for males $e^{.407(12)-.118(27)-1.514-1.749}/(1 + e^{.407(12)-.118(27)-1.514-1.749}) = .173$, for example, and the relative risks for males and females differ by a factor of about 3 (2.8:1 for females:males or .35:1 for males:females), not by a factor of 4 or 5. Suggestion: Do the math. There is no excuse here for approximations that can so easily be misleading.

3.5. More on Categorical Predictors: Contrasts and Interpretation

The use of zeros and ones to represent the different possible values of the variable ETHN in Figure 3.1 and Table 3.1 is called *indicator* coding because it indicates the presence or absence of a categorical attribute. Indicator coding is only one of several ways to treat design variables in logistic regression analysis. One alternative is called *simple* coding in SPSS LOGISTIC REGRESSION. With simple contrasts, logistic regression coefficients for the design variables are identical to the coefficients produced with indicator coding; only the intercept changes.

Another alternative is *deviation* coding, the default option in SPSS LOGISTIC REGRESSION. With deviation coding, the effect of each design variable is compared with the overall effect of the independent variable. This is analogous to comparing the means (not weighted by number of cases) for the three categories in regression or analysis of variance. In logistic regression, the deviation coding measures the deviation of the logit for each group from the average logit for the entire sample. With deviation coding, the reference category no longer has an arbitrary coefficient of 0. Instead, its coefficient is equal to the negative of the sum of the coefficients for the other categories. If computer time is more expensive than human labor, calculation of the omitted coefficient by hand may be reasonable. In a personal computer or free computing environment, however, it makes more sense to calculate two models, with different reference categories, to obtain not only the estimates for the coefficients of all three categories, but also the standard error and statistical significance of the otherwise omitted category.[17] The model with deviation coding for ETHN is summarized in Table 3.2.

Other than the changes in the individual coefficients for ETHN, the use of a different coding scheme for the indicator variables has not changed the results of the analysis. The order of the ethnic groups is the same as in Figure 3.1 (non-Hispanic European Americans have the lowest probability, and others the highest probability of marijuana use), but unlike Table 3.1, Table 3.2 adds the information that being African American would be, on the whole, *negatively* associated with the probability of marijuana use if the effect of ethnicity were statistically significant. The coefficients for indicator contrasts in Table 3.1 can be reconstructed from Table 3.2 by subtracting −.3388, the coefficient for non-Hispanic Europeans, from each of the other coefficients in Table 3.2. Deviation coding thus gives us information similar to indicator coding, but with a comparison to an "average" effect instead of a reference category.

TABLE 3.2

Logistic Regression for Prevalence of Marijuana Use: Deviation Coding of Ethnicity

Dependent Variable	*Association/ Predictive Efficiency*	*Independent Variable*	*Unstandardized Logistic Regression Coefficient (b)*	*Standard Error of b*	*Statistical Significance of b*	*Standardized Logistic Regression Coefficient*
PMRJ5	$G_M = 108.257$ $(p = .000)$	EDF5	.407	.069	.000	.531
		BELIEF4	−.118	.060	.048	−.143
	$R^2_L = .367$					
		SEX (male)	−1.514	.405	.000	−.233
	$R^2 = .435$					
		ETHN			.552	
	$\lambda_p = .488$	White	−.339	.319	.289	−.044
		Black	−.094	.391	.810	−.011
	$\tau_p = .604$	Other	.433	.388	.388	.033
		Intercept	−1.410	2.042	.490	—

Other contrasts available for logistic regression analysis in SPSS include Helmert, reverse Helmert, polynomial, repeated, and special contrasts. Helmert, reverse Helmert, orthogonal, and repeated contrasts are appropriate for testing whether the effects of different categories of an ordinal predictor are consistent with the ordering of the categories. Polynomial contrasts test for linear and nonlinear effects. The use of different contrasts for ordinal variables has no effect on the model fit or on the statistical significance of the categorical ordinal variable. The results may, however, suggest an appropriate recoding of the variable to take advantage of any apparent linearity, monotonicity, or any natural breaks between categories. The simplest ordinal contrast is the repeated (SPSS) or profile (SAS) contrast, in which each category of the independent variable except the first (the reference category) is compared to the previous category. By examining the coefficients for the categories, it is possible to see whether a monotonic or linear relationship exists between the independent variable and the dependent variable. If there is a nonsystematic pattern of positive and negative coefficients, a nonlinear, nonmonotonic relationship is indicated, and the independent variable is best treated as though it were nominal rather than ordinal.

Whenever design variables are used to represent the effect of a single *nominal* variable, it is important that the design variables be treated as a group, rather than as individual variables. The statistical significance of the individual design variables should be considered *only* if the design variables *as a group* have a statistically significant effect on the dependent variable. The statistical significance of the individual design variables should be interpreted as whether the effect of being in a certain category is statistically significantly different from being in the reference category (for indicator coding) or from the average effect of the categorical variable (in deviation coding), *given that the categorical variable has a statistically significant effect to begin with*. In SPSS, a test of the effect of the statistical significance of the nominal variable (all the design variables taken together) is provided. In SAS, a similar test can be obtained by comparing the model with and without the nominal variable, that is, with and without *all* of the indicator variables used to represent ethnicity, using stepwise procedures. For ordinal contrasts, the overall statistical significance of the design variable indicates only whether the categorical variable, *treated as a nominal variable,* has a statistically significant effect on the dependent variable. For ordinal contrasts, the statistical significance of the individual coefficients may provide important information about the form of the relationship between the categorical predictor and the dependent variable, even when the categorical variable does not appear to have a statistically significant effect on the dependent variable.

3.6. Interaction Effects

In some statistical software, we need to specify only the interaction term to be included and the software calculates the interaction term, includes it in the equation, and provides information about its statistical significance and the strength of its relationship to the dependent variable. In other software packages, it is necessary to separately calculate the interaction term (or terms, if the interaction involves a nominal variable with more than two categories) and add it (or them) to the model. The only complication here is when the interaction involves a nominal variable with more than two categories, in which case it may be necessary to compare the model with and without *all* of the interaction terms to determine whether the interaction is statistically and substantively significant. In linear regression, a conservative estimate of the statistical significance of the interaction effect is the statistical significance of the change in R^2 that results from adding the interaction effect to the model, and substantive significance is best evaluated as the magnitude of the change in R^2 (how much does the interaction add to our ability to predict the dependent variable?). In logistic regression, the corresponding criteria are the statistical significance of the change in G_M and the magnitude of the change in R^2_L.

In Table 3.3, two interaction terms are added to the model from Table 3.2. The interaction terms, which represent the interaction between SEX and EDF5 and the interaction between SEX and BELIEF4, test whether the effects of belief that it is wrong to violate the law and exposure to delinquent friends are different for males and females. This is a test for differences in the partial slopes of the curves that represent the relationship of PMRJ5 to EDF5 and BELIEF4 for males and females. Individually, neither of the interaction terms is statistically significant. Addition of the two interaction terms together results in a marginally significant change in G_M (4.656, $p = .098$), a small increase in R^2_L (.016), and *decreases* in $\lambda_p(-.025)$

TABLE 3.3

Testing for the Interactions of Sex With Belief and Exposure

Dependent Variable	*Association/ Predictive Efficiency*	*Independent Variable*	*Unstandardized Logistic Regression Coefficient (b)*	*Standard Error of b*	*Statistical Significance of b*	*Standardized Logistic Regression Coefficient*
PMRJ5	$G_M = 112.913$ ($p = .000$)	EDF5	.549	.126	.000	.662
		BELIEF4	−.161	.088	.067	−.180
	$R^2_L = .383$ (R^2_L change $= .016$)	SEX (male)	−1.516	.408	.000	−.215
		ETHN			.459	
	$R^2 = .435$	Black	−.303	.516	.558	−.032
	(no change)	Other	.891	.761	.242	.063
	$\lambda_p = .463$ (change $= -.025$)	SEX × Z_{EDF}	−.919	.638	.150	−.187
	$\tau_p = .585$	SEX × Z_{BELIEF}	.451	.491	.358	.084
	(change $= -.019$)	Intercept	−2.132	3.122	.495	—

NOTE: EDF5 and BELIEF4 were standardized for inclusion in the interaction terms to avoid collinearity.

and $\tau_p(-.019)$ because there are two more false negatives in the prediction table when the interaction terms are added. The most reasonable conclusion from this result would be that the effects of belief that it is wrong to violate the law and exposure to delinquent friends have the same effects on the prevalence of marijuana use for males and females.

3.7. Stepwise Logistic Regression

The term "stepwise regression" could, in principle, be applied to the analysis of interaction effects in Table 3.3 insofar as that analysis involved a two-step procedure for testing whether interaction terms were appropriate in the model for prevalence of marijuana use. More often, the term refers to the use of decisions made by computer algorithms, rather than choices made directly by the researcher, to select a set of predictors for inclusion or removal from a linear or logistic regression model. Some authors defend stepwise techniques in this latter sense as a useful tool for exploratory research (Agresti & Finlay, 1997, pp. 527–534; Hosmer & Lemeshow, 1989, p. 106); others criticize it as an admission of ignorance about the phenomenon being studied (Studenmund & Cassidy, 1987). Without going too deeply into the arguments about the use of stepwise procedures, there appears to be general agreement that the use of computer-controlled stepwise procedures to select variables is inappropriate for theory testing because it capitalizes on random variations in the data, and produces results that tend to be idiosyncratic and difficult to replicate in any sample other than the sample in which they originally were obtained.

Proponents of the use of stepwise procedures suggest that they may be useful in two contexts: purely predictive research and exploratory research. In purely predictive research, there is no concern with causality, only with identifying a model, including a set of predictors, that provides accurate predictions of some phenomenon. For example, a college admissions office may want to know what variables are good predictors of college success, not for theoretical development, but purely for the practical purpose of selecting students likely to succeed in college. In exploratory research, there may be a concern with theory construction and development to predict and explain a phenomenon, when the phenomenon is so new or so little studied that existing "theory" amounts to little more than empirically unsupported hunches about explanations for the phenomenon. An example of the use of stepwise techniques in exploratory research is provided by Wofford, Elliott, and Menard (1994).

Wofford et al. studied the continuity of domestic violence in a national probability sample of young men and women, 18 to 27 years old. Twenty-six predictors, based on the domestic violence literature, were included in their analysis. As part of the study, respondents who had reported being victims or perpetrators of domestic violence in 1984 were reinterviewed in 1987 to see whether the domestic violence had continued or been suspended since the 1984 interview. A total of 108 women (out of 807 in the original sample) reported being victims of domestic violence in 1983 and were reinterviewed in 1986. Wofford et al. constructed a logistic regression model that included all 26 predictors. Because theory in this area was not well developed and because the number of cases was small relative to the number of explanatory variables suggested in the literature, stepwise logistic regression was used.

Backward elimination rather than forward inclusion was selected as the method of stepwise regression. In some cases, a variable may appear to have a statistically significant effect only when another variable is controlled or held constant. This is called a *suppressor* effect (Agresti & Finlay, 1997, p. 368). One disadvantage to forward inclusion as a method for stepwise regression is the possible exclusion of variables involved in suppressor effects. With backward elimination, because both variables will already be in the model, there is less risk of failing to find a relationship when one exists. Usually, the results of backward elimination and forward inclusion methods of stepwise linear regression will produce the same results, but when the results differ, backward elimination may uncover relationships missed by forward inclusion.

To further prevent the failure to find a relationship when one exists, the usual .05 criterion for statistical significance probably should be relaxed. Based on their studies of forward stepwise regression, Bendel and Afifi (1977) suggested that .05 is too low and often excludes important variables from the model. Instead, they recommended that the statistical significance criterion for inclusion be set in a range from .15 to .20. This results in an increased risk of rejecting the null hypothesis when it is true (finding a relationship that is not really there), but a lower risk of failing to reject the null hypothesis when it is false (not finding a relationship that really is there). In exploratory research, as opposed to theory testing, there tends to be a greater emphasis on finding good predictors than on eliminating bad ones. Wofford et al. examined three models: a full model with all of the variables in the logistic regression equation, a reduced model with all variables for which $p > .10$ were eliminated (in practice, this was the same as using a .15 or .20 cutoff), and a further reduced model with all variables for which $p > .05$ eliminated. Table 3.4 presents the results.

The first part of Table 3.4 compares the three models. For the full model, G_{M} is not statistically significant, indicating that the predictor variables contribute no more than chance to the explanation of the dependent variable. Part of the reason for the failure of the model χ^2 to attain statistical significance is the small sample size; another may be the large number of variables included in the model. Model 2 ($p < .10$) has a smaller, but statistically significant G_{M}, as does model 3 ($p < .05$), with only one predictor in the model. The change in the model χ^2 (or equivalently, the change in D_{M}) from model 1 to model 2 and from model 1 to model 3 is not statistically significant. However, the change in G_{M} from model 2 to model 3 is statistically significant at the .01 level. For the full model, R_{L}^2 is .20, for model 2, it decreases to .15, and for model 3 it is only .03. Model 1 has a τ_{p} of .48; τ_{p} actually increases to .51 for model 2, but for model 3 it is only .14.

Model 2 was selected for further analysis because (1) G_{M} was statistically significant for the reduced models but not the full model, (2) model 2 provided a statistically significantly better fit than model 3, and did not fit statistically significantly worse than the full model, and (3) the changes in R_{L}^2 and τ_{p} were relatively small (and in opposite directions) for model 1 compared to model 2, but R_{L}^2 and τ_{p} were much lower for model 3 than for model 2. The results of model 2 are presented in the second half of Table 3.4. Substantively, they indicate that women who are welfare recipients, from a *higher* social class background, who have committed minor assaults but who have *not* committed felony assaults, who have *not* witnessed parental violence, who have experienced higher frequencies of serious violence in the relationship, and who have sought professional assistance are more likely to experience continuity rather than suspension of domestic violence. Full discussion of the substantive results can be found in Wofford et al. (1994).

Several methodological points regarding stepwise logistic regression are illustrated in Table 3.4. Probably the most important methodological point is that these results must be regarded as very tentative and

TABLE 3.4

Continuity of Marital Violence Victimization (Women)

$N = 108$	*Model 1: All Variables Included*	*Model 2: Maximum $p = .100$*	*Model 3: Maximum $p = .050$*
Model χ^2 G_{M} (degrees of freedom)	30.254 (28 df)	21.284 (7 df)	4.472 (1 df)
Statistical significance of G_{M}	.351	.003	.034
D_{M}	119.429	128.298	145.211
Change in G_{M} from previous model (degrees of freedom)	—	8.870 (21 df)	16.812 (6 df)
Statistical significance of change in G_{M} from previous model	—	.99	.010
R^2_{L}	.202	.150	.030
τ_{p}	.481	.509	.145

Individual Predictor Results for Model 2

Independent Variables	*b*	*Standard Error*	*p (Based on Likelihood Ratio Statistic)*
Welfare recipient	1.88	.95	.03
Social class background	−0.03	.02	.05
Prior minor assault	1.24	.53	.02
Prior felony assault	−1.07	. 62	.08
Witnessed parental violence	−1.70	.64	.00
Frequency of serious violence in relationship	0.12	.06	.05
Sought professional assistance	0.88	.53	.09

inconclusive. This is a search for plausible predictors, not a convincing test of any theory. Second, an important element of the stepwise procedure is the comparison of the full and reduced models. As suggested by Bendel and Afifi (1977), the .05 criterion for inclusion appears to be too severe; based on the comparisons of goodness-of-fit and predictive efficiency statistics, more reasonable results are obtained with a more liberal cutoff point for statistical significance. Third and finally, the variables identified in model 2 are good candidates for use in the prediction of domestic violence, but some may as easily be effects (for example, seeking professional assistance) as causes. Further development and testing of theory may be based on these results, but would require replication with other data and explanation (preferably in the form of a clear theoretical justification) of why these variables appear as predictors of continuity of domestic violence.

4. AN INTRODUCTION TO LOGISTIC REGRESSION DIAGNOSTICS

When the assumptions of logistic regression analysis are violated, calculation of a logistic regression model may result in one of three problematic effects: biased coefficients, inefficient estimates, or invalid statistical inferences. *Bias* refers to the existence of a systematic tendency for the estimated logistic regression coefficients to be too high or too low, too far from 0, or too close to 0, compared to the true values of the coefficients. *Inefficiency* refers to the tendency of the coefficients to have large standard errors relative to the size of the coefficient. This makes it more difficult to reject the null hypothesis (the hypothesis that there is no relationship between the dependent variable and the independent variable) even when the null hypothesis is false. *Invalid statistical inference* refers to the situation in which the calculated statistical significance of the logistic regression coefficients is inaccurate. In addition, *high leverage* cases, cases with unusually high or low values on the independent variables (not on the dependent variable, which has only two values), or *outliers* with unusual values on the dependent variable, given the values of the independent variables, may be *influential cases that* exert a disproportionate influence on the estimated parameters. This chapter focuses on the consequences of violations of logistic regression assumptions, and methods for detecting and correcting violations of logistic regression assumptions. Also considered here are methods for detecting outliers, high leverage cases, and influential cases in logistic regression, and alternative approaches to dealing with those cases.

4.1. Specification Error

The first and most important assumption in both linear and logistic regression analysis is that the model is correctly specified. Correct specification has two components: the functional form of the model is correct, and the model includes all relevant independent variables and no irrelevant independent variables. Misspecification may result in biased logistic regression coefficients, coefficients that are systematically overestimated or underestimated. In Chapter 1, we saw that application of the linear regression model to a dichotomous dependent variable appeared to be misspecified. This led us to examine the logistic regression model. It may nonetheless be the case that the logistic regression model, with logit(Y) as the dependent variable and with a *linear* combination of independent variables, is incorrect in its functional form. First, logit(Y) may be equal to a *nonlinear* combination of the independent variables. Second, the relationship among some or all of the independent variables may be *multiplicative* or *interactive,* rather than additive.

Misspecification as a result of using the logistic function, as opposed to a different S-shaped function, is less likely to be a problem. Aldrich and Nelson (1984) demonstrated that logit models (based on the logistic distribution) and probit models (based on the normal distribution) produce highly similar results. Hosmer and Lemeshow (1989, p. 168) noted that logistic regression models are highly flexible and produce very similar results to other models in the range of probabilities between .2 and .8. There is usually little theoretical basis for preferring an alternative model.

4.1.1. Omitting Relevant Variables and Including Irrelevant Variables

Including one or more irrelevant variables has the effect of increasing the standard error of the parameter estimates, that is, of reducing the efficiency of the estimates, without biasing the coefficients. The degree to which the standard errors are inflated depends on the magnitude of the correlation between the irrelevant included variable and the other variables in the model. If the irrelevant included variable is completely uncorrelated with the other variables in the equation, the standard errors may not be inflated at all, but this condition is extremely unlikely in practice.

Omitting relevant variables from the equation in logistic regression results in biased coefficients for the independent variables, to the extent that the omitted variable is correlated with the independent variables in the logistic regression equation. As in linear regression (Berry & Feldman, 1985), the direction of the bias depends on the parameter for the excluded variable, the direction of the effect of the excluded variable on the dependent variable, and the direction of the relationship between excluded and included variables. The magnitude of the bias depends on the strength of the relationship between the included and excluded variables. If the excluded variable is completely uncorrelated with the included variables, the coefficients may be unbiased, but in practice this is unlikely to occur. Bias is generally regarded as a more serious problem than inefficiency, but a small amount of bias may be preferable to massive inefficiency.

Omitted variable bias may occur because available theories have failed to identify all of the relevant predictors or causes of a dependent variable, or because theoretically relevant variables have been omitted. The pattern characteristic of omitted variable bias may also occur if the functional form of the model is misspecified. A linear specification of a nonlinear model may be computationally equivalent to the omission of a variable that represents a nonlinear component of the relationship between the dependent variable and an independent variable. An additive specification of a nonadditive model may be equivalent to the omission of a variable, specifically a variable constructed as the interaction of two other variables, from the model. When the omitted variable is neither a nonlinear term nor an interaction term, only theory (or perhaps a disappointingly low R^2_L) offers much hope of identifying and remedying the problem. When the excluded variable is really a nonlinear term or an interaction term, a function of variables already in the equation, the detection and correction of the problem can be considerably easier.

4.1.2. Nonlinearity in the Logit

In a linear regression model, the change in the dependent variable associated with a one-unit change in the independent variable is constant, equal to the regression coefficient for the independent variable. If the change in Y for a one-unit change in X depends on the value of X (as it does when Y is a dichotomous variable), the relationship is nonlinear. Correspondingly, when logit(Y) is the dependent variable, if the change in logit(Y) for a one-unit change in X is constant and does not depend on the value of X, we say that the logistic regression model has a linear form or that the relationship is *linear in the logit*, and the change in logit(Y) for a one-unit change in X is equal to the logistic regression coefficient. If the relationship is not linear in the logit, the change in logit(Y) for a one-unit change in X is not constant, but depends on the value of X.

There are several possible techniques for detecting nonlinearity in the relationship between the dependent variable, logit(Y), and each of the independent variables (Hosmer & Lemeshow, 1989, pp. 88–91). One is to treat each of the independent variables as a categorical variable and use an orthogonal polynomial contrast to test for linear, quadratic, cubic, and higher order effects in either bivariate logistic regression or in a multiple logistic regression model. If the independent variable has a large number of categories (for example, 20), the standard errors tend to be large, and neither the linear nor any of the nonlinear effects may appear to be statistically significant, even when a statistically significant linear effect exists. A second possibility is to use the Box–Tidwell transformation described by Hosmer and Lemeshow (1989, p. 90). This involves adding a term of the form $(X)\ln(X)$ (X multiplied by the natural logarithm of X) to the equation. If the coefficient for this variable is statistically significant, there is evidence of nonlinearity in the relationship between logit(Y) and X. Hosmer and Lemeshow noted that this procedure is not sensitive to small departures from linearity. Additionally, this procedure does not specify the precise form of the nonlinearity. If the relationship is nonlinear, further investigation is necessary to determine the pattern of the nonlinearity.

A third procedure suggested by Hosmer and Lemeshow is to aggregate cases into groups defined by the values of the independent variable X, calculate the mean of the dependent variable Y for each group, then take the logit of the mean of Y for each group and plot it against the value of the independent variable. For each value i of the independent variable X, the mean of Y is the probability $P(Y = 1 \mid X = i)$. One problem with this procedure arises if, for any value of X, Y is always either 1 or 0. If it is, then we cannot calculate logit($\overline{Y}$), which would be equal to $\pm\infty$, either infinitely large or infinitely small. It may be possible to overcome this problem by grouping adjacent categories with similar but unequal probabilities. This could conceal some of the nonlinearity in the relationship, however. Another possible option would be to assign an arbitrarily large mean (for example, .99) to groups with a mean of 1 and an arbitrarily small mean (.01) to groups with a mean of 0 to implement this method. An important advantage to this method is that, like graphical techniques generally, it helps identify the pattern of the nonlinearity. In addition, examination of the plot may help identify cases with unusual values on the independent variable or combinations of values on the dependent and independent variable.

Table 4.1 presents the results of a Box–Tidwell test for nonlinearity. In the first part of Table 4.1, the two nonlinear terms BTEDF = (EDF5)ln(EDF5) and BTBEL = (BELIEF4)ln(BELIEF4) are added to the model. Taken together, the effects of the two nonlinear interaction terms are statistically significant (change in $G_{\mathrm{M}} = 15.066$ with 2 degrees of freedom; $p = .005$), but based on the G_X, the likelihood ratio statistic for each of the nonlinear terms, only BTBEL is statistically significant ($G_X = 9.932$ with 1 degree of freedom, $p = .002$ for BTBEL; $G_X = 2.226$ with 1 degree of freedom, $p = .136$ for BTEXP). When BTEXP is removed from the model, the coefficient for BTBEL is still statistically significant ($G_X = 12.839$, 1 degree of freedom, $p = .000$), and inclusion of BTBEL in the equation increases R_{L}^2 by .034 (3.4%).

Figure 4.1 shows why the relationship between PMRJ5 and BELIEF4 appears to be nonlinear. The mean of PMRJ5 was calculated for each value of BELIEF4, and because the mean of PMRJ5 was either 0 or 1 for several values of BELIEF4, values of 1 were recoded as .99 and values of 0 were recoded as .01. Next, the logit of each mean was taken and plotted in Figure 4.1 against the values of BELIEF4. In the lower left quadrant of the plot there are two outliers, respondents who had very weak mean beliefs that it is wrong to violate the law, but who report no marijuana use. Each mean, as it turns out, is based on a single case. Except for these two cases, the plot does not appear to depart substantially from linearity. The second half of Table 4.1 confirms this assessment. With the two cases deleted from the analysis, BTEDF and BTBEL, separately and in combination, have no statistically significant effect on the logit of PMRJ5. Whether these cases should be deleted or retained will receive further consideration subsequently.

4.1.3. Nonadditivity

Nonlinearity occurs when the change in the dependent variable associated with a one-unit change in the independent variable depends on the value of the independent variable. Nonadditivity occurs when the change in the dependent variable associated with a

(text continues on page 75)

TABLE 4.1
Box–Tidwell Tests for Nonlinearity

Dependent Variable	*Association/ Predictive Efficiency*	*Independent Variable*	*Unstandardized Logistic Regression Coefficient (b)*	*Standard Error of b*	*Statistical Significance of b (Based on Likelihood Ratio G_X)*	*Standardized Logistic Regression Coefficient*
PMRJ5	$G_M = 123.322$	EDF5	2.415	1.191	.076	3.073
	$(p = .000)$					
$n = 227$		BELIEF4	3.620	1.421	.002	–4.272
	$R^2_L = .418$					
		SEX (male)	–1.660	.428	.000	–.249
	Change in G_M					
	from base	ETHN			.516	
	model	Black	.333	.536	.535*	.036
	(Table 3.1)	Other	.859	.817	.293*	.063
	=15.065					
	$(p = .001)$	BTEDF	–.551	.325	.136	–1.620
	Change in R^2_L	BTBEL	–.891	.340	.002	–4.172
	from base					
	model = .051	Intercept	–29.210	8.733	.001*	—

PMRJ5	$G_{\text{M}} = 125.195$	EDF5	2.411	1.184	.076	3.012
	($p = .000$)					
$n = 225$		BELIEF4	1.380	2.812	.644	−1.503
	$R^2_{\text{L}} = .427$					
		SEX (male)	−1.676	.436	.000	−.249
	Change in G_{M}					
	from base	ETHN			.530	
	model	Black	.342	.534	.522*	.036
	($n = 225$)	Other	.823	.814	.312*	.060
	=2.561					
	($p = .278$)					
		BTEDF	−.555	.323	.137	−2.551
	Change in R^2_{L}					
	from base	BTBEL	−.371	.660	.599	−1.716
	model					
	($n = .225$)	Intercept	−15.541	17.134	.364*	—
	=.009					

*Statistical significance based on the likelihood ratio statistic is not available for individual categories of categorical independent variables or for the intercept; for these, the Wald statistic is used to determine statistical significance.

```
                              PLOT OF LMPMJB WITH BELIEF4
         +---+----+----+----+----+----+----+----+----+----+----+----+----+--+
        5+                                                                  +
         |                                                                  |
         |               1    1    1       1 1                              |
         |                                                                  |
         |                                                                  |
     3.75+                                                                  +
         |                                                                  |
         |                                                                  |
         |                                                                  |
         |                                                                  |
      2.5R                                                                  +
         |                                                                  |
         |                                                                  |
         |                                      1                           |
         |                                                                  |
     1.25+                                        1                         +
         |                              1                                   |
L        |                                                                  |
M        |                                                                  |
P        |                                                1                 |
M       0+                                                                  +
J        |                                             1    1               |
B        |                                           1                      |
         |                                                            1     R
         |                                                     1            |
    -1.25+                                                                  +
         |                                                       1          |
         |                                                          1       |
         |                                                                  |
         |                                                                  |
     -2.5+                                                                  +
         |                                                                  |
         |                                                                  |
         |                                                                  |
         |                                                               1  |
    -3.75+                                                                  +
         |                                                                  |
         |                                                                  |
         |   1                   1                                          |
         |                                                                  |
       -5+                                                                  +
         +---+----+----+----+----+----+----+----+----+----+----+----+----+--+
                 10        14        18        22        26        30
             8        12        16        20        24        28        32

                                      BELIEF4
```

Figure 4.1. Logistic Regression Diagnostics: Test for Nonlinearity
LMPMJB = Logit of Mean of Prevalence of Marijuana use for each value of BELIEF4. BELIEF4 = Belief that it is wrong to violate the law.

one-unit change in the independent variable depends on the value of one of the other independent variables. For example, a one-unit change in exposure to delinquent friends may produce a larger change in the frequency or prevalence of marijuana use for individuals with weak to moderate beliefs that it is wrong to violate the law (who may be more susceptible to peer influence) than in individuals who strongly believe that it is wrong to violate the law (who may be less susceptible to peer influence). Detection of nonadditivity is not as straightforward as detection of nonlinearity in either linear or logistic regression. Unless theory provides some guidance, we are commonly left with the choice between assuming an additive model, testing for interaction effects that seem intuitively plausible, or testing for all possible interaction effects. This last option is feasible for relatively simple models, but becomes progressively more tedious and carries increasingly more risk of capitalizing on random sampling variation as the number of variables in the model increases.

One example of an interaction effect was examined in Table 3.3. The results in Table 3.3 indicated that the effects of exposure and belief on the prevalence of marijuana use were not statistically significantly different for males and females. This is an example of an interaction between a continuous predictor (actually two: EDF5 and BELIEF4) and a dichotomous variable. Other possible patterns include interactions between two categorical predictors (SEX and ETHN) or between two continuous variables (an interaction between EDF5 and BELIEF4).

4.2. Collinearity

Collinearity (or colinearity or multicollinearity) is a problem that arises when independent variables are correlated with one another. *Perfect collinearity* means that an independent variable is a perfect linear combination of the other independent variables. If we treated each independent variable in turn as the dependent variable in a model with all of the other independent variables as predictors, perfect collinearity would result in an $R^2 = 1$ for each of the independent variables. When perfect collinearity exists, it is impossible to obtain a unique estimate of the regression coefficients; any of an infinite number of possible combinations of linear or logistic regression coefficients will work equally well. Perfect collinearity is rare, except as an oversight: the inclusion of three variables, one of which is the sum of the other two would be one example.

Less than perfect collinearity is fairly common. *Any* correlation among the independent variables is indicative of collinearity. As collinearity increases among the independent variables, linear and logistic regression coefficients will be unbiased and as efficient as they can be (given the relationships among the independent variables), but the standard errors for linear or logistic regression coefficients will tend to be large. More efficient unbiased estimates may not be possible, but the level of efficiency of the estimates may be poor. Low levels of collinearity are not generally problematic, but high levels of collinearity (perhaps corresponding to an $R^2 = .80$ or more for at least one of the independent variables) may pose problems, and very high levels of collinearity (perhaps corresponding to an $R^2 = .90$ or more for at least one of the independent variables) almost certainly result in coefficients that are not statistically significant, even though they may be quite large. Collinearity also tends to produce linear and logistic regression coefficients that appear to be unreasonably high: as a rough guideline, standardized logistic or linear regression coefficients greater than 1 or unstandardized logistic regression coefficients greater than 2 should be examined to determine whether collinearity is present.

For linear regression, detection of collinearity is straightforward. Most standard regression routines in widely used software packages provide optional information on the R^2 or some function of the R^2 for each of the independent variables, when it is treated as the dependent variable with all of the other independent variables as predictors. For example, the *tolerance* statistic, available in SAS PROC REG and in SPSS REGRESSION, is simply $1-R_X^2$, where R_X^2 is the variance in each independent variable, X, explained by all of the other independent variables. Corresponding to the rough guidelines outlined in the preceding text, a tolerance of less than .20 is cause for concern; a tolerance of less than .10 almost certainly indicates a serious collinearity problem. Although tolerance is not available in SAS PROC LOGISTIC or SPSS LOGISTIC REGRESSION, it can be obtained easily by calculating a linear regression model using the same dependent and independent variables that you are using in the logistic regression model. Because the concern is with the relationship among the *independent* variables, the functional form of the model for the dependent variable is irrelevant to the estimation of collinearity.

In Table 4.1, for both of the models with the nonlinear terms BTEDF and BTBEL, the logistic regression coefficients were somewhat high, and the standardized logistic regression coefficients for EDF5, BELIEF4, BTEDF, and BTBEL were all larger than 1. This suggests that there may be a problem of collinearity in the nonlinear model. Table 4.2 presents collinearity statistics, produced by an OLS regression routine, for two models. The first, labeled "Basic Model," is the logistic regression model from Table 3.1. The second, labeled "Nonlinear Model," is the model from the first half of Table 4.1 with the nonlinear terms BTEDF and BTBEL included. In both models, Black and Other are design variables for ETHN. For the basic model, all of the tolerances exceed .70, indicating no serious problem of collinearity. For the nonlinear model, Table 4.2 confirms what the standardized coefficients in Table 4.1 suggested: the tolerances for SEX and the two design variables for ETHN remain high, but BTEDF and BTBEL are severely collinear with EDF5 and BELIEF4, as indicated by tolerances less than .01.

The good news about collinearity is that it is easy to detect. The bad news is that there are few acceptable remedies for it. Deleting variables involved in collinearity runs the risk of omitted variable bias. Combining collinear variables into a single scale, for example, by factor analysis, suggests that the theory (if any) used to construct your model or the measurement process used to collect your data was faulty, casting doubt on any further inferences you may draw from your analysis. *Ridge regression* (Schaefer, 1986) allows the user to

TABLE 4.2
Testing for Collinearity

		Tolerance	
Dependent Variable	*Independent Variable*	*Basic Model*	*Nonlinear Model*
PMRJ5	EDF5	.717	.00249
	BELIEF4	.707	.00148
	SEX (male)	.994	.994
	ETHN		
	Black	.959	.958
	Other	.983	.974
	BTEDF	—	.00253
	BTBEL	—	.00147

produce somewhat more biased but substantially more efficient estimates by increasing the estimated variance of the variables (thereby decreasing the proportion of the variance that is explained). Perhaps the safest strategy is to focus on the combined effects of all of the variables in the model and to recognize the precariousness of any conclusions about individual predictors in the presence of high collinearity. For a more detailed discussion of remedies to collinearity, see Berry and Feldman (1985, pp. 46–50) or Fox (1991). Briefly, though, there is no really satisfactory solution to high collinearity.

4.3. Numerical Problems: Zero Cells and Complete Separation

When collinearity exists, it does not necessarily indicate that there is anything wrong with the model or the theory underlying the model. Instead, problems arise because of empirical patterns in the data (the high correlation among independent variables). Two related problems with similar symptoms are zero cell count and complete separation. Zero cell count occurs when the dependent variable is invariant for one or more values of a categorical independent variable. If, for example, all of the respondents in the Other category for ethnicity reported using marijuana (or if they all reported *not* using marijuana), we would have a problem with a zero cell in the contingency table for the relationship between prevalence of marijuana use and ethnicity. The odds of marijuana use for respondents other than white and black would be $1/(1-1) = 1-0 = +\infty$ and the logit = ln(odds) would also be $+\infty$, infinitely large. [If the prevalence of marijuana use were 0 for this group, the odds would be $0/(1-0) = 0$ and the logit would be $\ln(0) = -\infty$, infinitely small.] When the odds are 0 or 1 for a single individual or case, this is not a problem; when they are 0 or 1 for an entire group of cases, as defined by the value of a categorical independent variable, the result will be a very high estimated standard error for the coefficient associated with that category (including coefficients for which that category serves as a reference category).

The problem of zero cell count applies specifically to categorical variables, particularly nominal variables. For continuous variables and for ordinal categorical variables, it is common to have means of 0 or 1 for some values of the independent variables. The reason that this is not a problem is that we assume a certain pattern to the relationship between the dependent variable and the continuous predictor (linear in linear regression, logistic in logistic regression), and use that pattern to "fill in the blanks" in the distribution of the dependent variable over the values of the independent variable. For categorical variables, we are unable to assume such a pattern. Instead, when we find problems of zero cell count for categorical predictors, we must choose among (1) accepting the high standard errors and the uncertainty about the values of the logistic regression coefficients, (2) recoding the categorical independent variable in a meaningful way (either by collapsing categories or by eliminating the problem category) to eliminate the problem of zero cell count, or (3) adding a constant to each cell of the contingency table to eliminate zero cells.

The first option may be acceptable if we are concerned more with the overall relationship between a set of predictors and a dependent variable than with the effects of the individual predictors. The overall fit of the model should be unaffected by the zero cell count. The third option has no serious drawbacks, but Hosmer and Lemeshow (1989, p. 127) suggested that it may not be adequate for complex analyses. The second option results in cruder measurement of the independent variable, and may bias the strength of the relationship between the predictor and the dependent variable toward zero. However, if there is a conceptual link between some categories of the independent variable and if the distribution of the dependent variable across those categories appears similar, this may be a reasonable option. Usually, this will be done during univariate and bivariate screening of the data. A hidden example of this has been followed throughout this monograph to this point: the coding of ethnicity. In the original survey, ethnicity was divided into six categories: non-Hispanic European American, African American, Hispanic American, Native American, Asian American, and other. The last four categories were collapsed into a single category, Other, because of the small number of cases. Had they been retained in their original form, problems of zero cell count would have plagued the analyses.

If you are too successful in predicting the dependent variable with a set of predictors, you have the problem of complete separation. Both the logistic regression coefficients and their standard errors will tend to be extremely large. The dependent variable will be perfectly predicted: $G_M = D_0$, $D_M = 0$, $R_L^2 = 1$. If separation is less than complete (sometimes called *quasicomplete* separation), logistic regression coefficients and their standard errors still will be extremely large. An example of quasicomplete separation is given in Figure 4.2, based on artificially constructed data. If complete separation occurs in a bivariate logistic relationship, the logistic regression model cannot be calculated. Although there is nothing intrinsically wrong with complete separation (after all, perfect prediction is what we are trying to achieve), as a practical matter it should arouse our suspicions, becasue it almost never occurs in real-world research. Complete or quasi-complete separation may instead indicate problems in the data or the analysis, for example, having almost as many variables as there are cases to be analyzed.

Collinearity, zero cells, and complete separation have the common symptom of very large standard errors and often, but not always, large coefficients as well (Hosmer & Lemeshow, 1989). All, therefore, result in inefficient estimation of the parameters in the model. None, however, is known to result in biased parameters or in inaccurate (as opposed to inefficient) inferences. Problems with zero cell counts can be averted by careful univariate and bivariate analysis before logistic regression is used. Complete separation may indicate either an error that needs to be corrected or a brilliant breakthrough in theory and analysis. (Congratulations!) Most likely, it indicates a problem. Collinearity is the most bothersome of the three problems, because it indicates either a flaw in the theory, a flaw in the operationalization of the theory, or a problem in the empirical data that confounds the testing of the theory, insofar as the theory is concerned with the effects of individual predictors rather than with the combined effect of a set of predictors. Like zero cell counts, collinearity can be detected (with the help of a good multiple regression package) before logistic regression analysis begins. What to do about it if it is detected is problematic, more art than science.

4.4. Analysis of Residuals[18]

In linear regression, the residual is commonly denoted e, and $e_j = Y_j - \widehat{Y}_j$ is the difference between the observed and predicted values of Y for a given case j. This should be distinguished from the error of prediction, denoted ϵ_j, which represents the difference between the true value of Y_j in the population (a value that may be different from the observed value of Y in the sample, for example, as a result of measurement error) and the estimated value of Y_j, $\widehat{Y}_j$ (Berry, 1993). In linear regression, certain assumptions about errors (zero mean, constant variance or homoscedasticity, normal distribution, no correlation of error terms with one another, no correlation of error terms

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	13.003	.654	.872

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	42.448	1	.000
	Block	42.448	1	.000
	Model	42.448	1	.000

Classification Table(a)

			Predicted		
			TRUE		Percentage Correct
	Observed		0	1	
Step 1	TRUE	0	19	1	95.0
		1	1	19	95.0
	Overall Percentage				95.0

a The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	P3	219.720	74.535	8.690	1	.003	2.65E+95
	Constant	-109.860	37.275	8.687	1	.003	.000

a Variable(s) entered on step 1: P3.

```
          Step number: 1

          Observed Groups and Predicted Probabilities

       16 +                                                            +
          |                                                            |
          |                                                            |
 F        |                                                            |
 R     12 +                                                            +
 E        |                                                            |
 Q        |0     1                                              1     1|
 U        |0     0                                              1     1|
 E      8 +0     0                                              1     1+
 N        |0     0                                              1     1|
 C        |0     0                                              1     1|
 Y        |0     0                                              1     1|
        4 +0     0                                              1     1+
          |0     0                                              1     1|
          |0     0                                              1     1|
          |0     0                                              0     1|
Predicted ---------------+--------------+--------------+----------------
  Prob:   0             .25            .5             .75             1
  Group:  000000000000000000000000000000111111111111111111111111111111

          Predicted Probability is of Membership for 1
          The Cut Value is .50
          Symbols: 0 - 0
                   1 - 1
          Each Symbol Represents 1 Case.
```

Figure 4.2. Quasicomplete Separation

with independent variables) are necessary if we are to draw statistical inferences from a sample to a larger population. These assumptions may sometimes be tested by using the residuals, e_j as estimates of the errors ϵ_j. Violations of some assumptions (zero mean, normal distribution) may have relatively minor consequences. Violation of others is more problematic. Heteroscedasticity inflates standard errors and renders tests of statistical significance inaccurate, and may itself be a symptom of nonadditivity or nonlinearity. Correlation between the independent variable and the error term generally indicates misspecification, the effects of which may include bias, inefficiency, or inaccurate statistical inference.

In linear regression, the residuals are straightforwardly computed from the regression equation. In logistic regression, several different residuals are available, corresponding to the different levels (probability, odds, logit) at which the analysis may be conceptualized. The principal purpose for which residuals analysis is used in logistic regression is to identify cases for which the model works poorly or cases that exert more than their share of influence on the estimated parameters of the model.

The difference between the observed and the predicted probability is $e_j = P(Y_j = 1) - P(Y_j = 1)$, where $P(Y_j = 1)$ is the estimated probability that $Y = 1$ based on the model. As Hosmer and Lemeshow explained, in linear regression, we can assume that the error is independent of the conditional mean of Y, but in logistic regression, the error variance is a function of the conditional mean. For this reason, residuals (estimates of error) are standardized by adjusting them for their standard errors. The Pearson (Hosmer & Lemeshow, 1989) or standardized (SPSS) or chi (SAS) residual is

$$r_j = z_j = \chi_j = \frac{P(Y_j = 1) - \widehat{P}(Y_j = 1)}{\sqrt{\widehat{P}(Y_j = 1)[1 - \widehat{P}(Y_j = 1)]}}.$$

This is just the difference between the observed and estimated probabilities divided by the binomial standard deviation of the estimated probability. For large samples, the standardized residual, hereafter z_j, should be normally distributed with a mean of 0 and a standard deviation of 1. Large positive or negative values of z_j indicate that the model fits a case j poorly. Because z_j should have a normal distribution, 95% of the cases should have values between −2 and +2, and 99% of the cases should have values between −2.5 and +2.5.

An alternative or supplement to the Pearson residual is the deviance residual, which is equal to $d_j = -2\ln$ (predicted probability of correct group). The deviance residual is the contribution of each case to D_M. Like z_j, d_j should have a normal distribution with a mean of 0 and a standard deviation of 1 for large samples. A third residual, the logit residual, is equal to the residual e_j divided by its variance (instead of its standard deviation, as in the standardized residual). This may be written

$$l_j = \frac{P(Y_j = 1) - \widehat{P}(Y_j = 1)}{\widehat{P}(Y_j = 1)[1 - \widehat{P}(Y_j = 1)]}.$$

4.4.1. Nonnormality of Residuals

In OLS regression, it is usually assumed that the errors are normally distributed. In small samples, if this assumption is violated, it renders statistical inference based on the regression equation (for example, the statistical significance of the regression coefficients) inaccurate. In large samples, inaccuracy of statistical inference is considered inconsequential because of results of the central limit theorem, which, briefly, indicates that the distribution of the regression coefficients in repeated sampling for large enough samples will approach a normal distribution with known mean (equal to the population mean) and variance. In logistic regression, the errors are *not* assumed to have a normal distribution. Instead, it is assumed that the distribution of the errors follows a binomial distribution, which approximates a normal distribution only for large samples. If the residuals are used to estimate the errors and if they are normally distributed (for a large sample), we can be more confident that our inferential statistics are correct, because normal (the distribution we are considering) and binomial (the assumed distribution) distributions are about the same for large samples. Contrary to the situation in linear regression analysis, however, if we find that the residuals are not normally distributed for small samples, we need not necessarily be concerned about the validity of our statistical inferences.

We can test for normality by plotting the standardized or deviance residuals against a normal curve, or in a normal probability plot; see, for example, SPSS (1999b). More importantly, we can use the standardized and deviance residuals to identify cases for which the model fits poorly, cases with positive or negative standardized or deviance residuals greater than 2 in absolute value.[19] This may help us identify not only cases for which the model fits poorly, but also cases that exert a disproportionately large influence on the estimates for the model parameters.

4.4.2. Detecting and Dealing With Influential Cases

Cases that potentially have a large influence on the parameters of the logistic regression model may be identified in part[20] by high values of the leverage statistic, or hat value, h_i. In linear regression, the leverage statistic is derived from the equation $\widehat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{jn}Y_k = \sum h_{ij}Y_i$, and it expresses the predicted value of Y for a case j as a function of the observed values of Y for case j and for all of the other cases as well (Fox, 1991). Each coefficient h_{ij} captures the influence of the observed variable Y_i on the predicted value $\widehat{Y}_j$. It can be shown that $h_{ii} = \sum (h_{ij})^2$, so if we designate $h_i = h_{ii}$, we have a measure of the overall influence of Y_i on the predicted values of Y for all of the cases in the sample. The leverage is similarly derived in logistic regression (Hosmer & Lemeshow, 1989, pp. 150–151), and it varies between 0 (no influence) and 1 (it completely determines the parameters in the model). In an equation with k independent variables (including each design variable as a separate variable) or, equivalently, in an equation in which there are k degrees of freedom associated with G_{M}, the sum of the values of h_i is equal to $k+1$ and the mean value of h_i, $\sum h_i/N = (k+1)/N$. Cases with hat values larger than $(k+1)/N$ are high leverage cases.

Other indices of the influence of an individual case include the change in the Pearson χ^2 statistic and the change in D_M attributable to deleting the case from the analysis. The change in the Pearson χ^2 attributable to deleting a case j is $\Delta\chi_j^2 = z_j^2/(1-h_j)$, where z_j is the standardized residual and h_j is the leverage statistic for case j. The change in D_{M} is equal to $\Delta D_j = d_j^2 - z_j^2 h_j/(1-h_j) = d_j^2 - h_j(\Delta\chi^2)$, where d_j is the deviance residual, z_j is the standardized residual, and h_j^2 is the leverage statistic for case j. Both ΔD_j and $\Delta_j\chi_j^2$ have a χ^2 distribution, and their values should be interpreted accordingly. Their respective square roots should have an approximately normal distribution. If $\sqrt{\Delta D_j}$ (the Studentized residual provided in SPSS LOGISTIC REGRESSION or C in SAS PROC LOGISTIC) or $\sqrt{\Delta\chi_j^2}$ is less than -2 or greater than $+2$, it indicates a case that may be poorly fit and deserves closer inspection. The quantity $z_j^2 h_j/(1-h_j)$ is itself an indicator of the overall change in regression estimates attributable to deleting an individual observation, and is available in SAS PROC LOGISTIC as the optional statistic CBAR and in SPSS LOGISTIC REGRESSION as Cook's distance. A standardized version of this measure may be obtained by dividing Cook's distance by $(1-h_j)$; $z_j^2 h_j/(1-h_j)^2 =$ dbeta, the standardized change in the regression coefficients attributable to the deletion of case j. The leverage statistic and the related statistics described previously are all summary indicators of the influence of a case on the estimation of the model parameters. More detailed information can be obtained by examining changes in individual coefficients that occur when a case is deleted. The change in the logistic regression coefficient is described as the DFBETA in both SPSS and SAS.

4.4.3. Outliers and Residual Plots

Table 4.3 presents the results of an analysis of residuals. Cases with $\sqrt{\Delta D_j}$ less than -2 or greater than $+2$ were selected for examination. The table includes the sequential number of the case, the observed and predicted values of the case, the Pearson (ZResid), Studentized (SResid), and deviance (Dev) residuals, the leverage (Lever), and the deleted residuals ΔD_j (DIFDEV), $\Delta\chi^2$ (DIFCHI), and dbeta (DBETA). Part A of Table 4.3 presents the residuals for the model in Table 3.1. Part B presents the results with the most extreme outlier deleted. This case is one of the two identified in the analysis of nonlinearity in Figure 4.1. Part C presents the results with both of the outliers from Figure 4.1 deleted.

The first comment to be made about Table 4.3 is that several of the indicators are essentially redundant. The change in the Pearson χ^2 statistic, DIFCHI ($\Delta\chi^2$), is approximately equal to the Pearson residual, ZResid, squared. The deviance residual, Dev(d_j) is approximately equal to the Studentized residual (SResid), and the change in the deviance residual, DIFDEV (ΔD_j), is equal to the Studentized residual squared. The Pearson χ^2-based residuals are larger than the residuals based on D_{M} but they provide essentially the same information about the cases. The leverage and DBETA provide information that is not evident from the other diagnostics; it is similar but not redundant. Further analysis of the table focuses on the Pearson residual, the Studentized residual, the leverage, and DBETA.

TABLE 4.3

Logistic Regression Diagnostic Summaries

Case	*Observed PMRJ5*	*Pred*	*ZResid*	*Dev*	*SResid*	*Lever*	*DIFCHI*	*DIFDEV*	*DBETA*
				A. Full Model					
66	1	.0991	3.0143	2.1500	2.1637	.0127	9.20	4.68	.12
94	0	.8608	−2.4864	−1.9858	−2.0325	.0455	6.48	4.13	.31
139	1	.0815	3.3565	2.2391	2.2668	.0243	11.55	5.14	.29
148	1	.0612	3.9183	2.3641	2.3762	.0102	15.51	5.65	.16
178	0	.9914	−10.7055	−3.0823	−3.0983	.0103	115.80	9.60	1.21
201	1	.0650	3.7937	2.3383	2.3526	.0121	14.57	5.53	.18
$G_M = 108.257$		5 df	$p = .0000$	$R^2_L = .367$					
				B. Most Extreme Case Deleted					
1	0	.9122	−3.2239	−2.2059	−2.2557	.0436	10.87	5.09	.50
66	1	.0894	3.1913	2.1975	2.2116	.0127	10.32	4.89	.13
94	0	.8786	−2.6903	−2.0536	−2.0999	.0436	7.57	4.41	.34
139	1	.0861	3.2577	2.2146	2.2444	.0264	10.90	5.04	.30
148	1	.0510	4.3137	2.4396	2.4515	.0097	18.79	6.01	.18
200	1	.0661	3.7601	2.3312	2.3463	.0129	14.32	5.51	.19
$G_M = 118.156$		5 df	$p = .0000$	$R^2_L = .401$					
				C. Outliers From Figure 4.1 Deleted (Cases 178 and 1)					
66	1	.0808	3.3732	2.2432	2.2573	.0125	11.52	5.10	.15
94	0	.8858	−2.7852	−2.0832	−2.1289	.0425	8.10	4.53	.36
133	0	.8675	−2.5584	−2.0105	−2.0506	.0387	6.81	4.20	.27
139	1	.0885	3.2099	2.2023	2.2332	.0275	10.59	4.99	.30
148	1	.0441	4.6536	2.4982	2.5097	.0092	21.86	6.30	.20
200	1	.0675	3.7175	2.3221	2.3378	.0135	14.01	5.47	.19
$G_M = 122.634$		5 df	$p = .0000$	$R^2_L = .416$					

NOTE: Cases with Studentized residuals greater than 2.0000000 are listed.

In part A, case number 178 stands out. The Pearson residual is an enormous −10.7, the Studentized residual is greater than 3 in absolute value, and DBETA is greater than 1, all indicators of an extremely poor fit. Deleting this case would result in an improvement in G_M of 9.899 (1 degree of freedom, $p = .003$) and an increase in R^2_L of .034. Clearly the model would work better with this case deleted. In part B, with case 178 deleted, no case stands out as clearly. Case 148 has the highest Pearson residual and the highest Studentized residual, but would produce relatively little change in the logistic regression coefficients if deleted. Case 1 would have more of an effect on the logistic regression coefficients (DBETA = .50) and has the fourth highest Pearson residual and the third highest Studentized and deviance residuals of the six cases selected as outliers. Case 1 has the additional feature that it is one of the two outliers identified in Figure 4.1 as introducing nonlinearity into the model. With case 1 deleted, G_M improves by 4.478 (1 degree of freedom, $p = .038$) and R^2_L increases by .015. The improvement that results from deleting case 1 is considerably smaller than the improvement from removing case 178.

Should cases 1 and 178 be removed from the analysis? The answer to this question requires closer examination of the data. The two cases in question are both white, one male and one female. Both report low levels of belief that it is wrong to violate the law, but neither uses marijuana or hard drugs and both report very low levels of alcohol use as well. Case 1 (female) has a slightly higher level of belief and a substantially lower level of exposure to delinquent friends than case 178 (male), and is therefore less inconsistent with the model than case 178. Although unusual, the results are plausible and both cases probably should be retained. It would be useful, however, to extend the model to include variables that might explain why an individual who sees nothing wrong with breaking the law chooses not to use alcohol, marijuana, or other illicit drugs.

Landwehr, Pregibon, and Shoemaker (1984) and Hosmer and Lemeshow (1989) discussed graphical techniques for logistic regression diagnostics. These techniques offer a visual rather than numerical representation that may be more intuitively appealing to some researchers. For example, Hosmer and Lemeshow (1989) recommended plots of DIFCHI, DIFDEV, and DBETA with the predicted values to detect outlying cases. Examples of these plots are provided in the first column of plots in Figure 4.3. Each of the three plots represents two curves, one declining from left to right (cases for which the observed value of PMRJ5 is 1) and one increasing from left to right (cases for which the observed value of PMRJ5 is 0). Cases in the upper left and right corners of the plot are cases for which the model fits poorly. In the plot of the χ^2 change (DIFCHI) with the predicted value, one case is an extreme outlier, with DIFCHI greater than 100. From Table 4.3, we can see that this is case 178. Similarly, case 178 is the case in the plot of DIFDEV with a value of DIFDEV greater than 8 and in the plot of DBETA with a value greater than 1. On the scale of the plots in the first column of Figure 4.3, all of the other cases seem to cluster fairly close together. Once case 178 is deleted, however, the scale of the plots may be changed and other cases appear as outliers.

Full Model

```
          PLOT OF DIFCHI WITH PRED1PMJ
     ++----+----+----+----+----+----+----++
      |                                1  |
      |                                   |
   100+                                   +
D     |                                   |
I     |                                   |
F     |                                   |
C     |                                   |
H   50+                                   +
I     |                                   |
      |                                   |
      |  1                                |
      |  12 1                     21      |
     0+ RFFGCDA5B68454511423612315245723A +
     ++----+----+----+----+----+----+----++
          .15       .45       .75     1.05
      0        .3        .6        .9

          PLOT OF DIFDEV WITH PRED1PMJ
     ++----+----+----+----+----+----+----++
      |                                1  |
      |                                   |
     8+                                   +
D     |                                   |
I     |                                   |
F     |  2                                |
D     |   2                               |
E    4+    11                     21      +
V     |      312                          |
      |        32411       2111           |
      |            125114 11              |
      |     874644332    223 12152        |
     0+ RFFGC4                    45723A  +
     ++----+----+----+----+----+----+----++
          .15       .45       .75     1.05
      0        .3        .6        .9

           PLOT OF DBETA WITH PRED1PMJ
     ++----+----+----+----+----+----+----++
      |                                1  |
      |                                   |
     1+                                   +
D     |                                   |
B     |                                   |
E     |                                   |
T     |                                   |
A   .5+                                   +
      |        1                  1       |
      |   1                1      11      |
      |  2 11  1    1         1           |
      |   1  313211212 13  11             |
     0+ RFFGCC7464733231 1234 2215245723A +
     ++----+----+----+----+----+----+----++
          .15       .45       .75     1.05
      0        .3        .6        .9

               Predicted Value

              227 cases plotted.
```

Model with case 178 deleted

```
            PLOT OF DIFCHI WITH PRED1PMJ
    ++----+----+----+----+----+----+----++
     |  1                                |
     |                                   |
   16+                                   +
     |  1                                |
     |                                   |
     |   1                         1     |
     |   1                               |
    8+                            1      +
     |                           1       |
     |     121                   1       |
     |       2331           1            |
     |          13121422 13111           |
    0+ XGLAAB27543 16    12412 22416574B +
    ++----+----+----+----+----+----+----++
         .15       .45       .75     1.05
     0        .3        .6        .9

         PLOT OF DIFDEV WITH PRED1PMJ
    ++----+----+----+----+----+----+----++
     |  1                                |
     |  1                                |
    5+   2                         1     +
     |                            1      |
     |     1                     1       |
     |      211                  1       |
     |       221                         |
  2.5+         221         12            +
     |          212 1    31              |
     |             1222                  |
     |         23 161   22411            |
     |    9AB2752          1 224161      |
    0+ XGL1                        474B  +
    ++----+----+----+----+----+----+----++
         .15       .45       .75     1.05
     0        .3        .6        .9

           PLOT OF DBETA WITH PRED1PMJ
    ++----+----+----+----+----+----+----++
     |                                   |
     |                             1     |
  .45+        1                          +
     |                                   |
     |                            1      |
     |   1                       1       |
     |                   1               |
 .225+     1         1           1       +
     |  2   1  1    11      1            |
     |   1   1  11       1 1             |
     |      112    1    1                |
     |      111312122212 22 1   1        |
    0+ XGLAAA26443  4 1  11312 22316574B +
    ++----+----+----+----+----+----+----++
         .15       .45       .75     1.05
     0        .3        .6        .9

               Predicted Value

              226 cases plotted.
```

Figure 4.3. Logistic Regression Diagnostic Plots

The second column of plots in Figure 4.3 presents the same plots with case 178 deleted. The plots may appear to be more spread out than the plots in the first column, but this is only because the scale has changed (from 0–100 to 0–20 for DIFCHI, from 0–10 to 0–6.25 for DIFDEV, and from 0–1.25 to 0–.5 for DBETA). The combination of the two curves for PMRJ5 = 1 and PMRJ5 = 0 takes on a characteristic goblet-shaped pattern (especially for DIFDEV), with the outliers again located in the upper right and left corners of the plot, and also in the "cup" of the goblet. In contrast to the first column, there is a relatively smooth transition from the upper corners to the rest of the graph. The most outlying cases are not as sharply separated as case 178 was in the first column. This is reflected numerically in Table 4.3, part B, in which none of the Studentized residuals is larger than 2.5 in absolute value and none of the DBETAs is greater than 1.

4.5. Overdispersion and Underdispersion[21]

Logistic regression assumes binomial errors, and thus it is assumed that the variance $\sigma_Y^2 = P_{Y=1}(1 - P_{Y=1})$. For casewise data, when there is only one case per covariate pattern, this condition is satisfied. For grouped data, however, when cases are aggregated by covariate pattern and the covariate patterns are treated as the "cases," this assumption may be violated for one of several reasons, including the omission of an important predictor, clustering within the sample, or simply because the underlying distribution of the population is different from the assumed distribution. Overdispersion refers to the case in which σ_Y^2 is larger than expected, and underdispersion refers to the case in which σ_Y^2 is less than expected. For grouped data, if we calculate $\delta = \mathrm{D}/\mathrm{df}$, where D is the deviance statistic calculated by covariate pattern (as reported, for example, in the goodness-of-fit table in SPSS NOMREG) and df is the degrees of freedom associated with the deviance statistic (and reported in the same table), then *overdispersion* is indicated by $\delta > 1$ and *underdispersion* is indicated by $\delta < 1$. Underdispersion and overdispersion result in incorrect standard errors and thus incorrect inferential statistics, but it is possible to adjust the standard error by multiplying by the square root of δ: adjusted standard error = standard error × $\sqrt{\delta}$. SPSS NOMREG offers the option of correcting the dispersion using either the Pearson or the deviance χ^2, or by using a user-specified value for δ (N in SPSS NOMREG). Note in Figure 3.2 that the deviance $\chi^2 = 158.266$ with 151 df, so

$\delta = 158.266/151 = 1.05$, suggesting little or no overdispersion or underdispersion.

4.6. A Suggested Protocol for Logistic Regression Diagnostics

Testing for collinearity should be a standard part of any logistic regression analysis. It is quick, simple to implement with existing regression software, and may provide valuable information about potential problems in the logistic regression analysis before the analysis is undertaken. The Box–Tidwell test for nonlinearity is quick, easy to perform, and not overly sensitive to minor deviations from linearity, and should also be incorporated as a standard procedure in logistic regression. Whether to test for nonadditivity by modeling interactions among the independent variables depends on whether there are theoretical or other reasons to believe that such interactions exist. Modeling nonlinearity and nonadditivity should be approached with some caution, however. There is a real danger of overfitting a model, building in components that really capture random variation, rather than systematic regularities in behavior.

Using logistic regression diagnostics, like using linear regression diagnostics, is more art than science. The diagnostic statistics hint at potential problems, but what those problems are and whether remedial action is required, can be decided only after closer inspection of the data for the unusual cases. In a sample of 200 to 250, random sampling variation alone will produce 10 to 12 cases with values greater than 2 or less than −2 on standardized, normally distributed variables such as the deviance residual or the Studentized residual. Even cases with very large residuals, like case 178 in Table 4.3, do not necessarily indicate problems in the model, insofar as we are dealing with nondeterministic models in which individual human choice and free will may naturally produce less than perfect prediction of human behavior.

As a general approach, it seems appropriate to perform at least a limited set of diagnostics on any model as a precaution against miscoded data and a guide to weaknesses in our conceptual models. A minimal set of diagnostics might include the Studentized residual, the leverage, and the dbeta. Studentized residuals less than −3 and greater than +3 definitely deserve closer inspection; values less than −2 or greater than +2 may also warrant some concern. The disadvantage to the Pearson residual is that the information it provides tends to duplicate the information provided by the deviance and Studentized residuals, and the deviance residual, not the Pearson residual, is the criterion for estimating the parameters of the model and is thus somewhat more pertinent to the analysis of residuals. The advantage to the Pearson residual is that, because it has larger values than the deviance residual, outlying cases sometimes stand out more sharply (as does case 178 in part A of Table 4.3) with the Pearson residual than with the deviance or Studentized residual. Leverage values several times the expected value of $(k + 1)/N$ (which was about $5/227 = .02$ in this example) also deserve close attention. Large values of dbeta, especially values greater than 1 (remember, this is a standardized measure), also deserve closer examination. Whether the information contained in these diagnostics is presented visually is a matter of taste. The critical concern is that extreme values on these diagnostics require closer inspection of the data, and possibly reconsideration of the model.

5. POLYTOMOUS LOGISTIC REGRESSION AND ALTERNATIVES TO LOGISTIC REGRESSION

Logistic regression analysis may be extended beyond the analysis of dichotomous variables to the analysis of categorical (nominal or ordinal) dependent variables with more than two categories. In the literature on logistic regression, the resulting models have been called polytomous, polychotomous, or multinomial logistic regression models. Here, the terms dichotomous and polytomous will be used to refer to logistic regression models, and the terms binomial and multinomial will be used to refer to logit models from which polytomous logistic regression models may be derived. For polytomous dependent variables, the logistic regression model may be calculated as a special case of the multinomial logit model (Agresti, 1990; Aldrich & Nelson, 1984; DeMaris, 1992; Knoke & Burke, 1980).

Mathematically, the extension of the dichotomous logistic regression model to polytomous dependent variables is straightforward. One value (typically the first or last) of the dependent variable is designated as the reference category, $Y = h_0$, and the probability of membership in other categories is compared to the probability of membership in the reference category. For nominal variables, this may be a direct comparison, like the indicator contrasts for independent variables in the logistic regression model for dichotomous variables. For an ordinal variable, contrasts may be made with successive categories, in a manner similar to repeated or Helmert contrasts for independent variables in dichotomous logistic regression models.

For dependent variables with some number of categories M, this requires the calculation of $M-1$ equations, one for each category relative to the reference category, to describe the relationship between the dependent variable and the independent variables. For each category of the dependent variable except the reference category, we may write the equation

$$g_h(X_1, X_2, \ldots, X_k) = e^{(a_h + b_{h1}X_1 + b_{h2}X_2 + \cdots + b_{hk}X_k)}, \qquad h = 1, 2, \ldots, M-1, \qquad [5.1]$$

where the subscript k refers, as usual, to specific independent variables X and the subscript h refers to specific values of the dependent variable Y. For the reference category, $g_0(X_1, X_2, \ldots, X_k) = 1$. The probability that Y is equal to any value h other than the excluded value h_0 is

$$P(Y = h|X_1, X_2, \ldots, X_k) = \frac{e^{(a_h + b_{h1}X_1 + b_{h2}X_2 + \cdots + b_{hk}X_k)}}{1 + \sum_{h=1}^{M-1} e^{(a_h + b_{h1}X_1 + b_{h2}X_2 + \cdots + b_{hk}X_k)}}, \qquad h = 1, 2, \ldots, M-1, \qquad [5.2]$$

and for the excluded category $h_0 = M$ or 0,

$$P(Y = h_0|X_1, X_2, \ldots, X_k) = \frac{1}{1 + \sum_{h=1}^{M-1} e^{(a_h + b_{h1}X_1 + b_{h2}X_2 + \cdots + b_{hk}X_k)}} \qquad h = 1, 2, \ldots, M-1. \qquad [5.3]$$

Note that when $M = 2$, we have the logistic regression model for the dichotomous dependent variable, the reference category is the first category, $h_0 = 0$, and we have a total of $M - 1 = 1$ equations to describe the relationship. Logistic regression models for polytomous nominal dependent variables can be calculated in SAS using

CATMOD and in SPSS prior to version 10 using LOGLINEAR, both general log-linear analysis routines in which the calculation of polytomous logistic regression models is rather cumbersome. In SPSS as of version 10, however, NOMREG provides a more user-friendly approach to logistic regression models for nominal dependent variables. SAS LOGISTIC and SPSS PLUM provide similarly user-friendly routines for ordered polytomous dependent variables. Although the focus of this monograph is on SAS and SPSS, it is also worth noting that STATA (1999) provides a broad range of routines for logistic regression, including MLOGIT for nominal dependent variables and OLOGIT for ordinal dependent variables.

To illustrate the use of polytomous logistic regression, the dependent variable from previous examples, prevalence of marijuana use, is replaced by drug user type. Drug user type has four categories.

1. Nonusers report that they have not used alcohol, marijuana, heroin, cocaine, amphetamines, barbiturates, or hallucinogens in the past year.
2. Alcohol users report having used alcohol, but no illicit drugs, in the past year.
3. Marijuana users report having used marijuana (and, except in one case, using alcohol as well).
4. Polydrug users report illicit use of one or more of the "hard" drugs (heroin, cocaine, amphetamines, barbiturates, hallucinogens). Polydrug users also report using alcohol and, except in one case (a respondent who reported a single incident of hard drug use), marijuana as well.

The four categories can reasonably be regarded as being ordered from least serious to most serious drugs, in terms of legal consequences. Alternatively, with respect to the nonlegal consequences of the drugs, the scale could arguably be regarded as nominal. Both ordinal and nominal models of this variable will be considered. One additional change is made from previous models. Because the dependent variable has four categories and because of the small number of cases in the category "other" on the variable ethnicity (ETHN), ethnicity was recoded into two categories, white and nonwhite, for the following analyses. Failure to do this would have resulted in problems with zero cells, and instability in estimates of coefficients and their standard errors.

5.1. Polytomous Nominal Dependent Variables

Figure 5.1 presents the output from SPSS NOMREG[22] with DRGTYPE as a dependent variable, using a contrast for DRGTYPE that compares, in succession, (a) nonusers with alcohol users, (b) nonusers with marijuana users, and (c) nonusers with polydrug users. The resulting functions, $g_1(X)$, $g_2(X)$, and $g_3(X)$ may be defined as

$$\begin{aligned} g_1 &= \text{logit (probability of using some alcohol versus nonuse} \\ &\qquad \text{of drugs),} \end{aligned}$$

$$g_2 = \text{logit (probability of using marijuana versus nonuse of drugs),}$$

and

$$\begin{aligned} g_3 &= \text{logit (probability of using other illicit drugs versus nonuse} \\ &\qquad \text{of drugs).} \end{aligned}$$

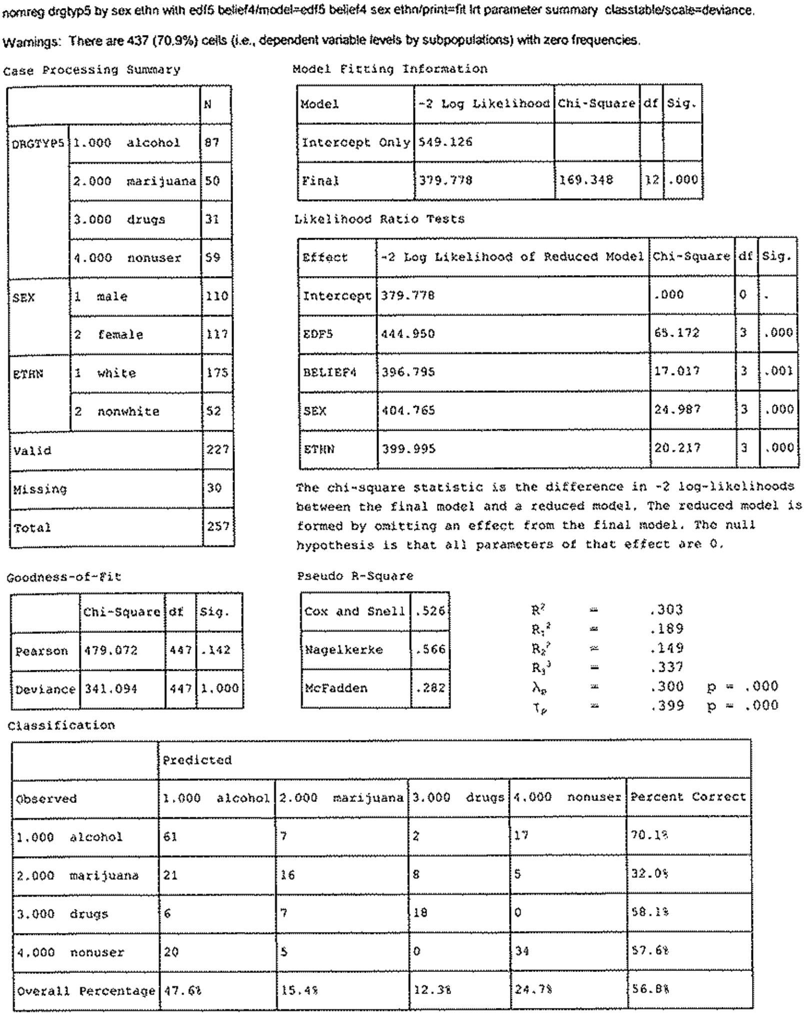
The equations for g_1, g_2, and g_3 using unstandardized coefficients are, from Figure 5.1,

$$\begin{aligned} g_1 &= .165(\text{EDF5}) - .271(\text{BELIEF4}) + .505(\text{SEX}) \\ &\quad + 1.616(\text{WHITE}) + 5.085, \\ g_2 &= .506(\text{EDF5}) - .285(\text{BELIEF4}) - .920(\text{SEX}) \\ &\quad + .357(\text{WHITE}) + 2.503, \end{aligned}$$

and

$$\begin{aligned} g_3 &= .633(\text{EDF5}) - .360(\text{BELIEF4}) - 2.224(\text{SEX}) \\ &\quad + 2.209(\text{WHITE}) + .768. \end{aligned}$$

The calculation of R^2 or η^2 and the standardized logistic regression coefficients is done separately for each logistic function, g_1, g_2, and g_3. (This is similar to the calculation of separate canonical correlation coefficients and standardized discriminant function coefficients for each linear discriminant function in discriminant analysis; see Klecka, 1980.) R^2 for the full model is calculated based on the predicted probabilities and observed classification for all four categories. Prediction tables are included in SPSS NOMREG, and can



```
nomreg drgtyp5 by sex ethn with edf5 belief4/model=edf5 belief4 sex ethn/print=fit lrt parameter summary  classtable/scale=deviance.
```

Warnings: There are 437 (70.9%) cells (i.e., dependent variable levels by subpopulations) with zero frequencies.

Case Processing Summary

		N
DRGTYP5	1.000 alcohol	87
	2.000 marijuana	50
	3.000 drugs	31
	4.000 nonuser	59
SEX	1 male	110
	2 female	117
ETHN	1 white	175
	2 nonwhite	52
Valid		227
Missing		30
Total		257

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	549.126			
Final	379.778	169.348	12	.000

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	379.778	.000	0	.
EDF5	444.950	65.172	3	.000
BELIEF4	396.795	17.017	3	.001
SEX	404.765	24.987	3	.000
ETHN	399.995	20.217	3	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	479.072	447	.142
Deviance	341.094	447	1.000

Pseudo R-Square

Cox and Snell	.526
Nagelkerke	.566
McFadden	.282

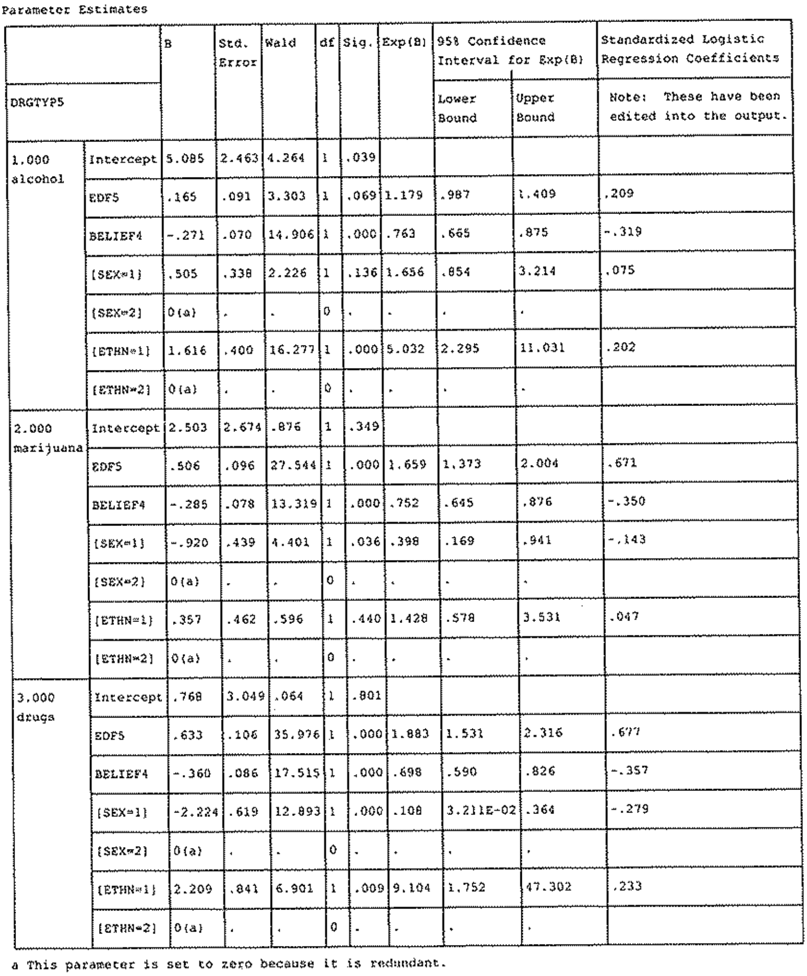
R^2 = .303
R_1^2 = .189
R_2^2 = .149
R_3^3 = .337
λ_p = .300 p = .000
τ_p = .399 p = .000

Classification

	Predicted				
Observed	1.000 alcohol	2.000 marijuana	3.000 drugs	4.000 nonuser	Percent Correct
1.000 alcohol	61	7	2	17	70.1%
2.000 marijuana	21	16	8	5	32.0%
3.000 drugs	6	7	18	0	58.1%
4.000 nonuser	20	5	0	34	57.6%
Overall Percentage	47.6%	15.4%	12.3%	24.7%	56.8%

Figure 5.1. Polytomous Nominal Logistic Regression

be constructed for SAS CATMOD by calculating the probability of classification for each value of *Y*, including the reference category, using Equations 5.2 and 5.3, then classifying each case into the category of *Y* for which it has the highest probability. The table itself can then be constructed using SAS PROC FREQ. Once the classi-



Parameter Estimates

		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)		Standardized Logistic Regression Coefficients
DRGTYP5								Lower Bound	Upper Bound	Note: These have been edited into the output.
1.000 alcohol	Intercept	5.085	2.463	4.264	1	.039				
	EDF5	.165	.091	3.303	1	.069	1.179	.987	1.409	.209
	BELIEF4	-.271	.070	14.906	1	.000	.763	.665	.875	-.319
	[SEX=1]	.505	.338	2.226	1	.136	1.656	.854	3.214	.075
	[SEX=2]	0(a)	.	.	0	.	.	.	.	
	[ETHN=1]	1.616	.400	16.277	1	.000	5.032	2.295	11.031	.202
	[ETHN=2]	0(a)	.	.	0	.	.	.	.	
2.000 marijuana	Intercept	2.503	2.674	.876	1	.349				
	EDF5	.506	.096	27.544	1	.000	1.659	1.373	2.004	.671
	BELIEF4	-.285	.078	13.319	1	.000	.752	.645	.876	-.350
	[SEX=1]	-.920	.439	4.401	1	.036	.398	.169	.941	-.143
	[SEX=2]	0(a)	.	.	0	.	.	.	.	
	[ETHN=1]	.357	.462	.596	1	.440	1.428	.578	3.531	.047
	[ETHN=2]	0(a)	.	.	0	.	.	.	.	
3.000 drugs	Intercept	.768	3.049	.064	1	.801				
	EDF5	.633	.106	35.976	1	.000	1.883	1.531	2.316	.677
	BELIEF4	-.360	.086	17.515	1	.000	.698	.590	.826	-.357
	[SEX=1]	-2.224	.619	12.893	1	.000	.108	3.211E-02	.364	-.279
	[SEX=2]	0(a)	.	.	0	.	.	.	.	
	[ETHN=1]	2.209	.841	6.901	1	.009	9.104	1.752	47.302	.233
	[ETHN=2]	0(a)	.	.	0	.	.	.	.	

a This parameter is set to zero because it is redundant.

Figure 5.1. (Continued)

fication table has been constructed, indices of predictive efficiency can be calculated as they have been for the Classification table in Figure 5.1, using the procedures described in Chapter 2. It is for polytomous models with nominal dependent variables that the differences between λ_p and τ_p, as opposed to other proposed indices of predictive efficiency, become most evident.

In Figure 5.1, the model works fairly well, as indicated by the statistically significant model χ^2 and the McFadden R_L^2 of .28. The explained variance in logit(Y) varies by the category of the dependent variable and is highest for g_3 (polydrug use) and lowest for g_2 (marijuana use). In the overall model, as indicated by the Likelihood Ratio Tests table, all four of the predictors are statistically significant. As indicated at the top of Figure 5.1 in the SPSS NOMREG statement, the dispersion has been corrected using the deviance χ^2 (/scale = deviance). This is because the deviance χ^2 appears to be somewhat lower than the degrees of freedom ($\chi^2 = 341$, df = 447, χ^2/df = .76), indicating underdispersion. The adjustment for dispersion will affect the statistical significance of the Wald coefficients. For alcohol use, the standardized coefficients (not part of the SPSS output, but like λ_p, τ_p, and R^2 added to the output) indicate that the best predictor is belief that it is wrong to violate the law, followed by ethnicity (white respondents are more likely to use alcohol than nonwhites). Exposure to delinquent friends is marginally significant according to the Wald statistic ($p = .069$), and gender is not statistically significant. For both marijuana and polydrug use, the best predictor is exposure to delinquent friends, followed by belief, then gender. Ethnicity is not a statistically significant predictor for marijuana use, but white respondents are more likely than nonwhite respondents to be polydrug users. Based on the Classification table in Figure 5.1, the indices of predictive efficiency $\lambda_p = .300$ and $\tau_p = .399$ are both statistically significant and moderately strong.

5.2. Polytomous or Multinomial Ordinal Dependent Variables

When the dependent variable is measured on an ordinal scale, many possibilities for analysis exist, including, but by no means limited to, logistic regression analysis. For a more detailed discussion, see Agresti (1990, pp. 318–332), Long (1997, pp. 114–147), or Clogg and Shihadeh (1994). Briefly, the options available include

1. Ignoring the ordering of the categories of the dependent variable and treating it as nominal
2. Treating the variable as though it were measured on a true ordinal scale
3. Treating the variable as though it were measured on an ordinal scale, but the ordinal scale represented crude measurement of an underlying interval/ratio scale
4. Treating the variable as though it were measured on an interval scale.

One possibility consistent with the first option is the use of a multinomial logit or logistic regression model for a nominal categorical dependent variable, as in Figure 5.1. Also possible under option 1 would be the use of discriminant analysis (Klecka, 1980). An example of the second option is the use of a *cumulative logit* model, in which the transformation of the dependent variable incorporates not only each category compared to a reference category, but also a comparison of each category with *all* of the categories with higher (or lower) numeric codes than the present category. The third option, assuming an underlying interval scale, could be implemented in LISREL by using weighted least squares (WLS) analysis of polychoric correlations (Jöreskog & Sörbom, 1988).[23] The fourth option might be implemented by using OLS regression with an ordinal dependent variable.

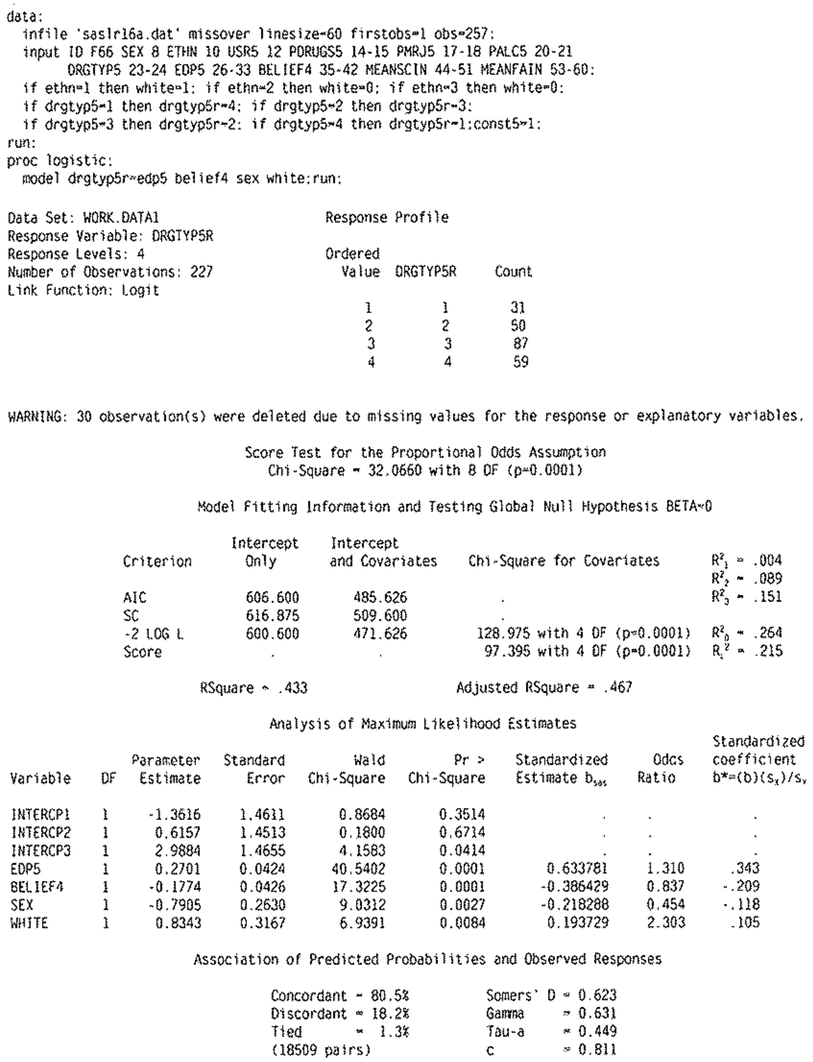
Selecting one of the options is a matter requiring careful judgment. The fourth option effectively assumes that the data are measured more precisely than they really are, but for ordinal variables with a large number of categories, it may be reasonable. The use of WLS with polychoric correlations appears to be a better option; it can be used with both large and small numbers of categories, and for most ordinal variables. The assumption of imprecise measurement of a quantity that is really continuous (political conservatism, seriousness of drug use) is inherently plausible. Both of these options allow predicted values that lie outside the range of observed values, but under the assumption of imprecise measurement, this may be reasonable.

Mechanical application of options available in existing software packages is *not* recommended. For example, SAS PROC LOGISTIC and SPSS PLUM can calculate polytomous logistic regression models for ordinal dependent variables, but both use a *cumulative logit model* for the dependent variable. This model assumes that the coefficient for each independent variable is invariant across the three equations, that is, $b_{\text{EDF5},1} = b_{\text{EDF5},2} = b_{\text{EDF5},3}$, $b_{\text{SEX},1} = b_{\text{SEX},2} = b_{\text{SEX},3}$, etc. (*parallel slopes*), where the variable in the subscript is the variable to which the coefficient refers, and the number in the subscript is the equation (1, 2, or 3) in which the coefficient appears. For the parallel slopes model, only the intercept is different for the three equations; otherwise, the effects of the independent variables are assumed to be constant across group comparisons. *It is important to emphasize that although this model is easily calculated using SAS PROC LOGISTIC or*

SPSS PLUM, it may not be the most appropriate model for the relationship between the dependent variable and the predictors.

Figure 5.2 summarizes the results of analyzing drug user type, DRGTYPE, as an ordinal variable in SAS PROC LOGISTIC. SAS provides a test of the assumption that the slopes are equal, the Score test. For Figure 5.2, the Score test of the null hypothesis that the slopes are equal is 32.066 with 8 degrees of freedom, statistically significant at the .0001 level. Because the Score test is statistically significant, the parallel slopes assumption is rejected, indicating that a model that does not assume parallel slopes would be more appropriate. The reasons for the rejection of the equal slopes model are evident from Figure 5.1: the variation in both the strength and statistical significance of the effects of EDF5 (not statistically significant for alcohol users as opposed to nonusers), SEX (not statistically significant for alcohol users as opposed to nonusers; stronger for polydrug users than for marijuana users as opposed to nonusers), and ETHN (not statistically significant for marijuana users as opposed to nonusers). The pattern of the differences in the coefficients in Figure 5.1 (especially the down-and-up pattern of the coefficients for ethnicity) suggests that treating DRGTYPE as a categorical nominal variable may be the best option.

SPSS PLUM provides much the same information as SAS LOGISTIC, except that SPSS PLUM excludes the information at the bottom of Figure 5.2 (Association of Predicted Probabilities and Observed Responses) and (as in SPSS NOMREG) includes the Pearson and deviance goodness-of-fit χ^2 statistics and the McFadden R^2_{L}, the latter of which (along with R^2 for the overall model and for each of the separate functions) has been edited into the SAS output in Figure 5.2. SPSS PLUM also offers alternatives to the logit distribution for dependent variables that are normally distributed, positively or negatively skewed, or have many extreme values. In both SPSS PLUM and SAS LOGISTIC, it is possible to save predicted values, and to use the predicted and observed values to produce contingency tables (in SAS PROC FREQ or SPSS CROSSTABS) to analyze the accuracy of classification. Doing so for Figure 5.2 would result in $\lambda_{\mathrm{p}} = .229$ ($p = .000$) and $\tau_{\mathrm{p}} = .208$ ($p = .000$), both smaller than in Figure 5.1, further suggesting that the dependent variable may better be treated as nominal rather than ordinal. For an ordinal variable in general, however, the statistics at the bottom of Figure 5.2, particularly the familiar ordinal measures of association Gamma and Tau-a,



```
data:
  infile 'saslr16a.dat' missover linesize=60 firstobs=1 obs=257:
  input ID F66 SEX 8 ETHN 10 USR5 12 PDRUGS5 14-15 PMRJ5 17-18 PALC5 20-21
       DRGTYP5 23-24 EDP5 26-33 BELIEF4 35-42 MEANSCIN 44-51 MEANFAIN 53-60:
  if ethn=1 then white=1: if ethn=2 then white=0: if ethn=3 then white=0:
  if drgtyp5=1 then drgtyp5r=4: if drgtyp5=2 then drgtyp5r=3:
  if drgtyp5=3 then drgtyp5r=2: if drgtyp5=4 then drgtyp5r=1:const5=1:
run:
proc logistic:
  model drgtyp5r=edp5 belief4 sex white:run:

Data Set: WORK.DATA1                      Response Profile
Response Variable: DRGTYP5R
Response Levels: 4                        Ordered
Number of Observations: 227                 Value  DRGTYP5R     Count
Link Function: Logit
                                              1        1          31
                                              2        2          50
                                              3        3          87
                                              4        4          59


WARNING: 30 observation(s) were deleted due to missing values for the response or explanatory variables.

                              Score Test for the Proportional Odds Assumption
                                 Chi-Square = 32.0660 with 8 DF (p=0.0001)

                         Model Fitting Information and Testing Global Null Hypothesis BETA=0

                            Intercept    Intercept
              Criterion       Only       and Covariates    Chi-Square for Covariates         R^2_1 = .004
                                                                                             R^2_2 = .089
              AIC            606.600      485.626                  .                         R^2_3 = .151
              SC             616.875      509.600                  .
              -2 LOG L       600.600      471.626         128.975 with 4 DF (p=0.0001)  R^2_0 = .264
              Score             .            .             97.395 with 4 DF (p=0.0001)  R_L^2 = .215

                         RSquare = .433                    Adjusted RSquare = .467

                                  Analysis of Maximum Likelihood Estimates
                                                                                             Standardized
                        Parameter   Standard       Wald          Pr >     Standardized     Odds   coefficient
Variable       DF      Estimate       Error   Chi-Square    Chi-Square   Estimate b_sas    Ratio   b*=(b)(s_x)/s_y

INTERCP1        1       -1.3616      1.4611       0.8684        0.3514                .       .         .
INTERCP2        1        0.6157      1.4513       0.1800        0.6714                .       .         .
INTERCP3        1        2.9884      1.4655       4.1583        0.0414                .       .         .
EDP5            1        0.2701      0.0424      40.5402        0.0001         0.633781   1.310       .343
BELIEF4         1       -0.1774      0.0426      17.3225        0.0001        -0.386429   0.837      -.209
SEX             1       -0.7905      0.2630       9.0312        0.0027        -0.218288   0.454      -.118
WHITE           1        0.8343      0.3167       6.9391        0.0084         0.193729   2.303       .105

                        Association of Predicted Probabilities and Observed Responses

                                Concordant = 80.5%         Somers' D = 0.623
                                Discordant = 18.2%         Gamma     = 0.631
                                Tied       =  1.3%         Tau-a     = 0.449
                                (18509 pairs)              c         = 0.811
```

Figure 5.2. SAS Output for Ordinal Logistic Regression

may be even more informative than λ_p or τ_p, because the former two measures, unlike the latter two, incorporate information on the ordering of the categories of the dependent variable.

5.3. Conclusion

The principal concern in using logistic regression analysis with polytomous dependent variables is not how to make the model work, but instead whether the logistic regression model is appropriate at all. For ordinal dependent variables, the problems that motivated the development of the logistic regression model (out-of-range predicted values of the dependent variable, heteroscedasticity) may not be present, and other models may be more appropriate than logistic regression, depending on assumptions about the underlying scale of the dependent variable and the functional form (linear, monotonic, nonmonotonic) of the relationship between the dependent variable and the independent variables. If there is an underlying interval scale, and if the relationships appear to be linear or monotonic, weighted least squares with polychoric correlations may be the best option. For nonmonotonic relationships, and especially when there are relatively few categories of the dependent variable, it may be best to treat the dependent variable as though it were nominal. When the dependent variable is nominal, or is an ordinal variable with few categories and is treated as though it were nominal, an alternative worth considering is discriminant analysis (Klecka, 1980). Another alternative, separate logistic regressions (Bess & Grey, 1984; Hosmer & Lemeshow, 1989, pp. 230–232), does not appear to produce results sufficiently consistent with the multinomial logit/polytomous logistic regression model to warrant its use. Only if polytomous logistic regression software were unavailable (an increasingly rare phenomenon, with all major software packages now including polytomous logistic regression routines) would this approach have any merit.

The ease of use, flexibility, broad applicability, and current popularity of logistic regression analysis make it particularly susceptible to misuse. Thoughtless and mechanical applications of logistic regression analysis will be no more fruitful than thoughtless and mechanical applications of linear regression or any other technique. It is important to recognize the weaknesses as well as the strengths of the method. Logistic regression is especially appropriate for the analysis of dichotomous and unordered nominal polytomous dependent variables. For ordinal polytomous dependent variables, it may be possible to use polytomous logistic regression analysis, but other models, including linear regression and weighted least squares with polychoric correlations, also deserve serious consideration. Polytomous ordinal variables are the dependent variables for which the technical motivation for using logistic regression is weakest and for which alternative methods of analysis are most likely to provide better solutions than logistic regression. Given these qualifications, however, the same ease of use (particularly improvements in logistic regression software, even in the few years since the first edition of this monograph), flexibility, and broad applicability of the logistic regression approach, as mentioned at the beginning of this paragraph, make logistic regression an extremely useful tool for analyzing a broad range of dependent variables for which OLS regression is not appropriate.

NOTES

1. Although the relationship being modeled often represents a causal relationship in which the single predicted variable is believed to be an effect of the one or more predictor variables, this is not always the case. We can as easily predict a cause from an effect (for example, predict whether different individuals are male or female based on their income) as predict an effect from a cause (predicting income based on whether someone is male or female). Throughout this monograph, the emphasis is on predictive rather than causal relationships, although the language of causal relationships is sometimes employed. Describing a variable as independent or dependent, therefore, or as an outcome or a predictor, does not necessarily imply a causal relationship. Instead, all relationships should be regarded as definitely predictive, but only possibly causal in nature.

2. Data are taken from the National Youth Survey, a national household probability sample of individuals who were adolescents (age 11 to 17) in 1976 and young adults (age 27 to 33) in 1992. Data were collected annually for the years 1976 to 1980, then in 3 year intervals thereafter, from 1983 to 1992. The data include information on self-reported illegal behavior, family relationships, school performance, and sociodemographic characteristics of the respondents. Details on sampling and measurement may be found in Elliott et al. (1985, 1989). For present purposes, attention is restricted to respondents who were 16 years old in 1980. In the scatterplot, the numbers and symbols refer to numbers of cases at a given point on the plot: a 1 indicates one case, a 2 indicates two cases, a 9 indicates nine cases; the letters A to Z continue the count, $A = 10$ cases, $B = 11$ cases, $\ldots, Z = 35$ cases. When more than 35 cases occupy a single point, an asterisk (*) is used.

3. For a review of levels of measurement, see, for example, Agresti and Finlay (1997, pp. 12–17).

4. The *unconditional mean* of Y is simply the familiar mean $\overline{Y} = \sum Y_j/N$. The *conditional mean* of Y for a given value of X is calculated by selecting only those cases for which X has a certain value and calculating the mean for those cases. The conditional mean can be denoted $\overline{Y}_{X=i} = \sum Y_{ij}/n_i$, where i is the value of X for which we are calculating the conditional mean of Y, Y_{ij} are the values of Y for the cases $(j = 1, 2, \ldots, n_i)$ for which $X = i$, and n_i is the number of cases for which $X = i$.

5. A brief discussion of probability, including conditional probabilities, is presented in the Appendix.

6. The logarithmic transformation is one of several possibilities discussed by Berry and Feldman (1985, pp. 63–64), Lewis-Beck (1980, p. 44), and others to deal with relationships that are nonlinear with respect to the variables, but may be expressed as linear relationships with respect to the parameters.

7. R_L^2 is also provided as the pseudo-R^2 in Stata (Stata, 1999). An earlier version of SAS PROC LOGISTIC [SAS (SUGI) PROC LOGIST (Harrell, 1986)] included a variant of R_L^2 that was adjusted for the number of parameters in the model. This

measure is analogous to the adjusted R^2 in linear regression, and we may denote it as R^2_{LA} to indicate its connection with R^2_L and to distinguish it from other R^2-type measures. $R^2_{LA} = (G_M - 2k)/(D_0)$, where k is the number of independent variables in the model. If $G_M < 2k$ and particularly if $G_M = 0$, it is possible to get a negative estimate for explained variance using R^2_{LA}.

8. It should be noted, however, that in other contexts, it may not be appropriate. For example, in proportional hazards models, R^2_L is more sensitive to censoring than R^2_N (Schemper, 1990, 1992).

9. The designation ϕ_p was selected because ϕ_p, like ϕ, is based on comparisons between observed and expected values for individual cells (rather than rows or columns, as with λ_p and τ_p), because the numerical value of ϕ_p is close to the numerical value of ϕ for tables with consistent marginals (row and column totals in which the larger row total corresponds to the larger column total), and ϕ_p and ϕ have the same sign (because they have the same numerator).

10. ϕ_p can be adjusted by adding the minimum number of errors, $|(a+b)-(a+c)| = |b-c|$, to the expected number of errors without the model. This results in a coefficient that (a) retains the proportional change in error interpretation (because the adjustment is built into the calculation of the expected error) and (b) still may have negative values if the model is pathologically inaccurate. For extremely poor models, the revised index still has a maximum value less than 1, even when the maximum number of cases is correctly classified, and the increment over ϕ_p is small. Based again on similarities with ϕ, we may designate this adjusted ϕ_p as ϕ'_p. Note, however, that ϕ'_p cannot be calculated as $\phi_p/\max(\phi_p)$; to do so would destroy the proportional change in error interpretation for the measure and would leave the measure undefined when the maximum value of ϕ_p was 0.

11. For a two-tailed test, the null hypothesis is that there is no difference between the proportion of errors with and without the prediction model. The alternative hypothesis is that the proportion of errors with the prediction model is not equal to the proportion of errors without the prediction model. For a one-tailed test, specifying that the model results in increased accuracy of prediction of the dependent variable, the null hypothesis is that the proportion of errors with the prediction model is no smaller than the proportion of errors without the prediction model. The alternative hypothesis is that the proportion of errors with the prediction model is less than the proportion of errors without the prediction model. If we want to know whether the prediction model improves our ability to predict the classification of the cases, the one-tailed test is more appropriate, and a negative value of λ_p will result in a negative value for d and failure to reject the null hypothesis.

12. Copas and Loeber (1990) noted this property and indicated that in this situation it would be a misinterpretation to regard a value of 1 as indicating perfect prediction. This leads to two questions. How *should* we interpret the value of RIOC in this situation? What value *does* indicate perfect prediction for RIOC? Ambiguity of interpretation is an undesirable quality in any measure of change, and there are enough better alternatives that the use of the RIOC measure should be avoided.

13. It will not always be the case that logistic regression produces a higher R^2 than linear regression for a dichotomous dependent variable. In a parallel analysis of theft for the full National Youth Survey sample, R^2 for linear and logistic regression was .255 and .253, respectively.

14. This is sometimes called a Type II error or a false negative (failure to detect a relationship that exists), as opposed to a Type I error or a false positive (concluding that there is a relationship when there really is none).

15. This was done using the backward stepwise procedure, to be discussed later in the text. In SPSS NOMREG and PLUM there is no stepwise procedure, but it is possible to include the likelihood statistics in the output.

16. If they were, they would indicate that non-Hispanic European Americans have the lowest rates of marijuana use, followed by African Americans, and others have the highest prevalence of marijuana use. It is always questionable, however, to make statements about the nature of a relationship that is not statistically significant and may reflect nothing more than random sample error.

17. Mathematically, the omitted category is redundant or of little or no interest. In both theory testing and applied research, however, it makes more sense to provide full information about the coefficients and their statistical significance for all three categories, rather than leave one for pencil and paper calculation.

18. Hosmer and Lemeshow (1989) distinguished between analyzing residuals based on individuals and analyzing residuals based on covariate patterns, the combinations of values of the independent variables that actually occur in the sample. When the number of covariate patterns is equal to the number of cases, or very nearly so, residuals must be analyzed for each case separately. This is the implicit approach taken in this section, and in SAS PROC LOGISTIC and SPSS LOGISTIC REGRESSION, but SPSS NOMREG aggregates cases by covariate pattern and produces the correct predictions, residuals, and goodness-of-fit tests based on those subpopulations (Norusis, 1999). When the number of cases is much larger than the number of covariate patterns or when some of the covariate patterns hold for more than five cases, Hosmer and Lemeshow recommended aggregating the cases by covariate pattern, because of potential underestimation of the leverage statistic h_j.

19. In a standard normal distribution with a mean of 0 and a standard deviation of 1, 95% of the cases should have standardized scores (or, in this context, standardized residuals) between −2 and +2, and 99% should have scores or residuals between −3 and +3. Having a standardized or deviance residual larger than 2 or 3 does not necessarily mean that there is something wrong with the model. We would expect about 5% of the sample to lie *outside* the range −2 to +2, and 1% to lie outside the range −2.5 to +2.5. Values far outside this range, however, are usually indications that the model fits poorly for a particular case and suggest either that there is something unusual about the case that merits further investigation or that the model may need to be modified to account for whatever it is that explains the poor fit for some of the cases.

20. As Fox (1991) noted, in linear regression, influence = leverage × discrepancy, where "discrepancy" refers to being an outlier on Y with respect to the predictors. In logistic regression, in contrast to linear regression, as fitted probabilities get close to 0 (less than .1) or 1 (greater than .9), the leverages stop increasing and turn rapidly toward 0 (Hosmer & Lemeshow, 1989, pp. 153–154).

21. A particularly clear and concise discussion of overdispersion and underdispersion may be found in Hutcheson and Sofroniou (1999). See also McCullagh and Nelder (1989, pp. 124–128).

22. Had this been run in SAS CATMOD, neither D_0 nor G_M would be directly available. For both, the likelihood χ^2 statistics are based on comparisons of cells in a contingency table, rather than on probabilities of category membership. The two are

related, however, and it is possible to derive the statistics appropriate for logistic regression analysis from the statistics provided by SAS PROC CATMOD. The appropriate steps are

1. Compute $D_0 = \sum(n_{Y=h})\ln[P(Y = h)] = \sum(n_{Y=h})\ln(n_{Y=h}/N)$, where $n_{Y=h}$ is the number of cases for which Y is equal to 1 of its possible values h, N is the total sample size, and the sum is taken over all possible values h of Y.
2. Examine the iteration history of the model. The "−2 Log Likelihood" from the final iteration, listed in the Maximum Likelihood Analysis table is approximately (but not exactly) equal to D_M.
3. Compute $G_M = D_0 - D_M$; compute $R_L^2 = G_M/D_0$.

If these procedures are followed with a dichotomous dependent variable, the resulting figures are *approximately* equal to the G_M and R_L^2 that would be obtained in the identical analysis from SAS PROC LOGISTIC. (SAS PROC LOGISTIC uses an iteratively reweighted least squares algorithm to calculate the model parameters; PROC CATMOD uses weighted least squares or maximum likelihood estimation, depending on the type of model being calculated.) The process is a bit awkward, but if the likelihood ratio statistics provided in CATMOD are used without modification, they produce results different from those that would be obtained using SAS PROC LOGISTIC when the dependent variable is dichotomous.

23. For reliable estimates, this requires a large N (Hu, Bentler, & Kano, 1992).

APPENDIX: PROBABILITIES

The probability of an event is estimated by its relative frequency in a population or sample. For example, if $n_Y = 1$ is the number of cases for which $Y = 1$ in a sample and N is the total number of cases in the sample, then

1. We denote the probability that Y is equal to 1 as $\mathrm{P}(Y=1)$
2. $\mathrm{P}(Y = 1) = n_{Y=1}/N$
3. The probability that Y is *not* equal to 1 is $\mathrm{P}(Y \neq 1) = 1 - \mathrm{P}(Y = 1) = 1 - (n_{Y=1}/N) = (N - n_{Y=1})/N$
4. The minimum possible value for a probability is 0 ($n_{Y=1} = 0$ implies $n_{Y=1}/N = 0$)
5. The maximum possible value for a probability is 1 ($n_{Y=1} = N$ implies $n_{Y=1}/N = 1$).

The *joint probability* of two independent events (occurrences that are unrelated to one another) is the product of their individual probabilities. For example, the probability that both X and Y are equal to 1, if X and Y are unrelated is $\mathrm{P}(Y = 1 \text{ and } X = 1) = \mathrm{P}(Y = 1) \times \mathrm{P}(X = 1)$. If X and Y are related (for example, if the probability that Y is equal to 1 depends on the value of X), then $\mathrm{P}(Y = 1$ and $X = 1)$ will not be equal to $\mathrm{P}(Y = 1) \times \mathrm{P}(X = 1)$. Instead, we will want to consider the *conditional probability* that $Y = 1$ when $X = 1$, or $\mathrm{P}(Y = 1|X = 1)$.

The conditional probability that $Y = 1$ is the probability that $Y = 1$ *for a given value of some other variable.* [In this context, we may sometimes refer to $\mathrm{P}(Y = 1)$, the probability that $Y = 1$ regardless of the value of any other variable, as the *unconditional* probability that $Y = 1$.] For example, the probability that the prevalence of marijuana use is equal to 1 for the data in Figure 2.1 is P(PMRJ5 = 1) =.35 (for males and females combined; detailed data not shown). The conditional probability that prevalence of marijuana use is equal to 1 is P(PMRJ5 = 1| SEX = 0) =.45 for females and P(PMRJ5 = 1 | SEX = 1) =.25 for males. For a dichotomous variable, coded as 0 or 1, the probability that the variable is equal to 1 is equal to the mean for that variable, and the conditional probability that the variable is equal to 1 is equal to the conditional mean (see note 4) for the variable.

REFERENCES

AGRESTI, A. (1990). *Categorical data analysis*. New York: Wiley.

AGRESTI, A., and FINLAY, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice–Hall.

ALDRICH, J. H., and NELSON, F. D. (1984). *Linear probability, logit, and probit models*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-045. Beverly Hills, CA: Sage.

ALLISON, P. D. (1999). *Logistic regression using the SAS system*. Cary, NC: SAS Institute.

BEGG, C. B., and GREY, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika, 71,* 11–18.

BENDEL, R. B., and AFIFI, A. A. (1977). Comparison of stopping rules in forward regression. *Journal of the American Statistical Association, 72,* 46–53.

BERRY, W. D. (1993). *Understanding regression assumptions*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-092. Newbury Park, CA: Sage.

BERRY, W. D., and FELDMAN, S. (1985). *Multiple regression in practice*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-050. Beverly Hills, CA: Sage.

BOHRNSTEDT, G. W., and KNOKE, D. (1994). *Statistics for social data analysis* (3rd ed.). Itasca, IL: F. E. Peacock.

BOLLEN, K. A. (1989). *Structural equation models with latent variables*. New York: Wiley.

BULMER, M. G. (1979). *Principles of statistics*. New York: Dover.

CLOGG, C. C., and SHIHADEH, E. S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage.

COPAS, J. B., and LOEBER, R. (1990). Relative improvement over chance (RIOC) for 2×2 tables. *British Journal of Mathematical and Statistical Psychology, 43,* 293–307.

COSTNER, H. L. (1965). Criteria for measures of association. *American Sociological Review, 30,* 341–353.

COX, D. R., and SNELL, E. J. (1989). *The analysis of binary data* (2nd ed.) London: Chapman and Hall.

CRAGG, J. G., and UHLER, R. (1970). The demand for automobiles. *Canadian Journal of Economics, 3,* 386–406.

DeMARIS, A. (1992). *Logit modeling*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-086. Newbury Park, CA: Sage.

ELIASON, S. R. (1993). *Maximum likelihood estimation: Logic and practice*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-096. Newbury Park, CA: Sage.

ELLIOTT, D. S, HUIZINGA, D., and AGETON, S. S. (1985). *Explaining delinquency and drug use*. Beverly Hills, CA: Sage.

ELLIOTT, D. S., HUIZINGA, D., and MENARD, S. (1989). *Multiple problem youth*. New York: Springer-Verlag.

FARRINGTON, D. P., and LOEBER, R. (1989). Relative improvement over chance (RIOC) and phi as measures of predictive efficiency and strength of association in 2×2 tables. *Journal of Quantitative Criminology, 5,* 201–213.

FOX, J. (1991). *Regression diagnostics.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-079. Newbury Park, CA: Sage.

HAGLE, T. M., and MITCHELL, G. E., II (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science, 36,* 762–784.

HARDY, M. (1993). *Regression with dummy variables.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-093. Newbury Park, CA: Sage.

HARRELL, F. E., Jr. (1986). The LOGIST procedure. In SAS Institute, Inc. (Ed.), *SUGI supplemental library user's guide,* (Version 5, pp. 269–293). Cary, NC: SAS Institute.

HOSMER, D. W., and LEMESHOW, S. (1989). *Applied logistic regression.* New York: Wiley.

HU, L., BENTLER, P. M., AND KANO, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351–362.

HUTCHESON, G., and SOFRONIOU, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models.* Thousand Oaks, CA: Sage.

JÖRESKOG, K. G., and SÖRBOM, D. (1988). *PRELIS: A program for multivariate data screening and data summarization* (2nd ed.). Chicago, IL: Scientific Software International.

JÖRESKOG, K. G., and SÖRBOM, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Chicago: Scientific Software International.

KLECKA, W. R. (1980). *Discriminant analysis.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-019. Beverly Hills, CA: Sage.

KNOKE, D., and BURKE, P. J. (1980). *Log-linear models.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-020. Beverly Hills, CA: Sage.

LANDWEHR, J. M., PREGIBON, D., and SHOEMAKER, A. C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association, 79,* 61–71.

LEWIS-BECK, M. S. (1980). *Applied regression: An introduction.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-022. Beverly Hills, CA: Sage.

LOEBER, R., and DISHION, T. (1983). Early predictors of male delinquency: A review. *Psychological Bulletin, 94,* 68–99.

LONG, J. S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage.

MADDALA, G. S. (1983). *Limited-dependent and qualitative variables in econometrics.* Cambridge, UK: Cambridge University Press.

MAGEE, L. (1990). R^2 measures based on Wøld and likelihood ratio joint significance tests. *The American Statistician, 44,* 250–253.

McCULLAGH, P. and NELDER, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall.

McFADDEN, D. (1974). The measurement of urban travel demand. *Journal of Public Economics, 3,* 303–328.

McKELVEY, R., and ZAVOINA, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4,* 103–120.

MENARD, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54,* 17–24.

MIECZKOWSKI, T. (1990). The accuracy of self-reported drug use: An evaluation and analysis of new data, In R. Weisheit (Ed.), *Drugs, crime, and the criminal justice system* (pp. 275–302). Cincinnati: Anderson.

NAGELKERKE, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78,* 691–692.

NORUSIS, M. J. (1999). *SPSS regression models 10.0.* Chicago: SPSS, Inc.

OHLIN, L. E., and DUNCAN, O. D. (1949). The efficiency of prediction in criminology. *American Journal of Sociology, 54,* 441–451.

SAS (1989). *SAS/STAT user's guide* (Version 6, 4th ed. Vols. 1 and 2). Cary, NC: SAS Institute.

SAS (1995). *Logistic regression examples using the SAS system.* Cary, NC: SAS Institute.

SCHAEFER, R. L. (1986). Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation, 25,* 75–91.

SCHEMPER, M. (1990). The explained variation in proportional hazards regression. *Biometrika, 77,* 216–218.

SCHEMPER, M. (1992). Further results on the explained variation in proportional hazards regression. *Biometrika, 79,* 202–204.

SCHROEDER, L. D., SJOQUIST, D. L., and STEPHAN, P. E. (1986). *Understanding regression analysis: An introductory guide.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-057. Beverly Hills, CA: Sage.

SIMONOFF, J. S. (1998). Logistic regression, categorical predictors, and goodness of fit: It depends on who you ask. *The American Statistician, 52,* 10–14.

SODERSTROM, I., and LEITNER, D. (1997). The effects of base rate, selection ratio, sample size, and reliability of predictors on predictive efficiency indices associated with logistic regression models. Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chicago.

SPSS (1991). *SPSS statistical algorithms* (2nd ed.). Chicago: SPSS, Inc.

SPSS (1999a). *SPSS advanced models 10.0.* Chicago: SPSS, Inc.

SPSS (1999b). *SPSS base 10.0 applications guide.* Chicago: SPSS, Inc.

STATA (1999). *Stata reference manual* (Release 6, Vol. 2). College Station, TX: Stata Press.

STUDENMUND, A. H., and CASSIDY, H. J. (1987). *Using econometrics: A practical guide.* Boston: Little, Brown.

VEALL, M. R., and ZIMMERMAN, K. F. (1996). Pseudo-R^2 measures for come common limited dependent variable models. *Journal of Economic Surveys, 10,* 241–260.

WIGGINS, J. S. (1973). *Personality and prediction: Principles of personality assessment.* Reading, MA: Addison–Wesley.

WOFFORD, S., ELLIOTT, D. S., and MENARD, S. (1994). Continuities in marital violence. *Journal of Family Violence, 9,* 195–225.

ABOUT THE AUTHOR

SCOTT MENARD is Research Associate in the Institute of Behavioral Science at the University of Colorado, Boulder. He received his A.B. at Cornell University and his Ph.D. at the University of Colorado, both in sociology. His primary substantive interests are in the longitudinal study of drug use and other forms of illegal behavior. His publications include the Sage QASS monograph *Longitudinal Research* (1991; second edition forthcoming in 2002) and the books *Perspectives on Publication* (with Elizabeth W. Moen, 1987), *Multiple Problem Youth* (with Delbert S. Elliott and David Huizinga, 1989), and *Juvenile Gangs* (with Herbert C. Covey and Robert J. Franzese, second edition 1997).